Laugh machine

Urbain, Jerôme ; Niewiadomski, Radoslaw ; Hofmann, Jennifer ; Bantegnie, Emeline ; Baur, Tobias ; Berthouze, Nadia ; Cakmak, Hüseyin ; Cruz, Richard Thomas ; Dupont, Stephane ; Geist, Matthieu ; Griffin, Harry ; Lingenfelser, Florian ; Mancini, Maurizio ; McKeown, Gary ; Miranda, Miguel ; Pammi, Sathish ; Pietquin, Olivier ; Piot, Bilal ; Platt, Tracey ; Ruch, Willibald ; Sharma, Abhishek ; Volpe, Gualtiero ; Wagner, Johannes

Abstract: The Laugh Machine project aims at endowing virtual agents with the capability to laugh naturally, at the right moment and with the correct intensity, when interacting with human participants. In this report we present the technical development and evaluation of such an agent in one specific scenario: watching TV along with a participant. The agent must be able to react to both, the video and the participant's behaviour. A full processing chain has been implemented, inte- grating components to sense the human behaviours, decide when and how to laugh and, finally, synthesize audiovisual laughter animations. The system was evaluated in its capability to enhance the affective experience of naive participants, with the help of pre and post-experiment questionnaires. Three interaction conditions have been compared: laughter-enabled or not, reacting to the participant's behaviour or not. Preliminary results (the number of experiments is currently to small to obtain statistically significant differences) show that the interactive, laughter-enabled agent is positively perceived and is increasing the emotional dimension of the experiment.

Proceedings eNTERFACE'12

The 8th International Summer Workshop on Multimodal Interfaces

July 2nd - July 27th 2012 ; Supélec, Metz, France

Prof. Olivier Pietquin, Chair

Supélec

Supélec
Metz Technopôle
2 rue Edouard Belin
57070 Metz, France

Ph.: +33 (0)3 87 76 47 70
Fax : +33 (0)3 87 76 47 00
Olivier.Pietquin@Supelec.fr
http://malis.metz.supelec.fr/~pietquin
http://enterface12.metz.supelec.fr

Forewords

The eNTERFACE'12 workshop was organized by the Metz' Campus of Supélec and co-sponsored by the ILHAIRE and Allegro European projects.

The previous workshops in Mons (Belgium), Dubrovnik (Croatia), Istanbul (Turkey), Paris (France), Genoa (Italy), Amsterdam (The Netherlands) and Plzen (Czech Republic) had an impressive success record and had proven the viability and usefulness of this original workshop. eNTERFACE'12 hosted by Supélec in Metz (France) took this line of fruitful collaboration one step further. Previous editions of eNTERFACE have already inspired competitive projects in the area of multimodal interfaces, has secured the contributions of leading professionals and has encouraged participation of a large number of graduate and undergraduate students.

We received high quality project proposals among which the 8 following projects were selected.

1. Speech, gaze and gesturing - multimodal conversational interaction with Nao robot
2. Laugh Machine
3. Human motion recognition based on videos
4. M2M -Socially Aware Many-to-Machine Communication
5. Is this guitar talking or what!?
6. CITYGATE, The multimodal cooperative intercity Window
7. Active Speech Modifications
8. ArmBand : Inverse Reinforcement Learning for a BCI driven robotic arm control

All the projects resulted in promising results and demonstrations which are reported in the rest of this document. The workshop gathered more than 70 attendees coming from 16 countries all around Europe and even further. We received 4 invited speakers (Laurent Bougrain, Thierry Dutoit, Kristiina Jokinen and Anton Batliner) whose talks were greatly appreciated. The workshop was held in a brand new 800 m2 building in which robotics materials as well as many sensors were available to the attendees. This is why we proposed a special focus of this edition on topics related to human-robot and human-environment interaction. This event was a unique opportunity for students and experts to meet and work together, and to foster the development of tomorrow's multimodal research community.

All this has been made possible thanks to the the good will of many of my colleagues who volunteered before and during the workshop. Especially, I want to address many thanks to Jérémy

who did a tremendous job for making this event as enjoyable and fruitful as possible. Thanks a lot to Matthieu, Thérèse, Danièle, Jean-Baptiste, Senthil, Lucie, Edouard, Bilal, Claudine, Patrick, Michel, Dorothée, Serge, Calogero, Yves, Eric, Véronique, Christian, Nathalie and Elisabeth. Organizing this workshop was a real pleasure for all of us and we hope we could make it a memorable moment of work and fun.

Olivier Pietquin

Chairman of eNTERFACE'12

The eNTERFACE'12 Sponsors

We want to express our gratitude to all the organizations which made this event possible.

The eNTERFACE'12 Scientific Committee

Niels Ole Bernsen, University of Southern Denmark - Odense, Denmark
Thierry Dutoit, Faculté Polytechnique de Mons, Belgium
Christine Guillemot, IRISA, Rennes, France
Richard Kitney, University College of London, United Kingdom
Benoît Macq, Université Catholique de Louvain, Louvain-la-Neuve, Belgium
Cornelius Malerczyk, Zentrum für Graphische Datenverarbeitung e.V, Germany
Ferran Marques, Univertat Politécnica de Catalunya PC, Spain
Laurence Nigay, Université Joseph Fourier, Grenoble, France
Olivier Pietquin, Supélec, Metz, France
Dimitrios Tzovaras, Informatics and Telematics Intsitute, Greece
Jean-Philippe Thiran, Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland
Jean Vanderdonckt, Université Catholique de Louvain, Louvain-la-Neuve, Belgium

The eNTERFACE'12 Local Organization Committee

General chair	Olivier Pietquin
Co-chair	Jeremy Fix
Web management	Claudine Mercier
Technical support	Jean-Baptiste Tavernier
Social activities	Matthieu Geist
Administration	Danielle Cebe
	Thérèse Pirrone

eNTERFACE 2012 - Project reports

Laugh Machine

Jérôme Urbain[1], Radoslaw Niewiadomski[2], Jennifer Hofmann[3], Emeline Bantegnie[4], Tobias Baur[5], Nadia Berthouze[6], Hüseyin Çakmak[1], Richard Thomas Cruz[7], Stéphane Dupont[1], Matthieu Geist[10], Harry Griffin[6], Florian Lingenfelser[5] Maurizio Mancini[8], Miguel Miranda[7], Gary McKeown[9], Sathish Pammi[2], Olivier Pietquin[10], Bilal Piot[10], Tracey Platt[3], Willibald Ruch[3], Abhishek Sharma[2], Gualtiero Volpe[8] and Johannes Wagner[5]

[1]TCTS Lab, Faculté Polytechnique, Université de Mons, Place du Parc 20, 7000 Mons, Belgium
[2]CNRS - LTCI UMR 5141 - Telecom ParisTech, Rue Dareau, 37-39, 75014 Paris, France
[3]Universität Zürich, Binzmuhlestrasse, 14/7, 8050 Zurich, Switzerland
[4]LA CANTOCHE PRODUCTION, Hauteville, 68, 75010 Paris, France
[5]Institut für Informatik, Universität Augsburg, Universitätsstr. 6a, 86159 Augsburg, Germany
[6]UCL Interaction Centre, University College London, Gower Street, London, WC1E 6BT, United Kingdom
[7]Center for Empathic Human-Computer Interactions, De la Salle University, Manila, Philippines
[8]Universita Degli Studi di Genova, Viale Francesco Causa, 13, 16145 Genova, Italy
[9]The Queen's University of Belfast, University Road, Lanyon Building, BT7 1NN Belfast, United Kingdom
[10]École Supérieure d'Électricité, Rue Edouard Belin, 2, 57340 Metz, France

***Abstract*—The Laugh Machine project aims at endowing virtual agents with the capability to laugh naturally, at the right moment and with the correct intensity, when interacting with human participants. In this report we present the technical development and evaluation of such an agent in one specific scenario: watching TV along with a participant. The agent must be able to react to both, the video and the participant's behaviour. A full processing chain has been implemented, integrating components to sense the human behaviours, decide when and how to laugh and, finally, synthesize audiovisual laughter animations. The system was evaluated in its capability to enhance the affective experience of naive participants, with the help of pre and post-experiment questionnaires. Three interaction conditions have been compared: laughter-enabled or not, reacting to the participant's behaviour or not. Preliminary results (the number of experiments is currently to small to obtain statistically significant differences) show that the interactive, laughter-enabled agent is positively perceived and is increasing the emotional dimension of the experiment.**

***Index Terms*—Laughter, virtual agent.**

I. Introduction

LAUGHTER is a significant feature of human communication, and machines acting in roles like companions or tutors should not be blind to it. So far, limited progress has been made towards allowing computer-based applications to deal with laughter. In consequence, only few interactive multimodal systems exist that use laughter in the interactions. Within the long term aim of building a truly interactive machine able to laugh and respond to human laughter, during the eNTERFACE Summer Workshop 2012 we have developed the Laugh Machine project.

This project had three main objectives. First of all we aimed to build an interactive system that is able to detect the human laughs and to laugh back appropriately (*i.e.*, right timing, right type of laughter) to the human and the context. Secondly, we wanted to use the laughing agent to support psychological studies investigating benefits of laughter in human-machine interaction and consequently improve the system towards more naturalness and believeability. The third aim was the collection of multimodal data on human interactions with the agent-based system.

To achieve these aims, we tuned and integrated several existing analysis components that can detect laughter events as well as interpreters that controlled how the virtual agent should react to them. In addition, we also provided output components that are able to synthesize audio-visual laughs. All these components were integrated to work in real-time. Secondly, we focused on building an interactive scenario where our laughing agent can be used. In our scenario, the participant watches a funny stimulus (*i.e.*, film clip, cartoon) together with the virtual agent. The agent is able to laugh, reacting to both, the stimulus and the user's behavior. We evaluated the impact of the agent through user evaluation questionnaires (*e.g.*, assessing the mood pre and post experiments, funniness and aversiveness ratings to both stimuli and agent behavior, etc.). At the same time we were able to collect multimodal data (audio, facial expressions, shoulder movements, and Kinect depth maps) of people interacting with the system.

This report is organized as follows. First, related work is presented in Section II. Then, the experimental scenarios are outlined in Section III, so that the framework for developing the technical setup is known. The data used for training the components is presented in section IV. Section V shows the global architecture of the Laugh Machine system. The next sections focus on the components of this system: details about the input components are given in Section VI, Section VII is related to the dialog manager and the output components are described in Section VIII. Then, the conducted experiments to evaluate the system are explained in Section IX. The results of

these experiments are discussed in Section X. Section XI refers to the data that has been collected during the experiments. Finally, Section XII presents the conclusions of the project.

II. Related work

Building an interactive laughing agent requires tools from several fields: at least audiovisual laughter synthesis for the output, and components able to detect particular events like participant's laughs and decide when and how to laugh. In the following paragraphs we will present the main works in audiovisual laughter recognition, acoustic laughter synthesis and visual laughter synthesis, then the interactive systems involving laughter that have already been built. Regarding a decision component dealing with laughter as input and output, to the best of our knowledge there is no existing work.

A. Audiovisual laughter recognition

In the last decade, several systems have been built to distinguish laughter from other sounds like speech. It started with audio-only detection. The global approach followed up to now for discriminating speech and laughter is to compute standard acoustic features (MFCCs, pitch, energy, ...) and feed them into typical classifiers: Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs) or Multi-Layer Perceptrons (MLPs). Kennedy and Ellis [1] obtained 87% of classification accuracy with SVMs fed with 6 MFCCs; Truong and van Leeuwen [2] reached slightly better results (equal error rate of 11%) with MLPs fed with Perceptual Linear Prediction features; Knox and Mirghafori [3] obtained better performance (around 5% of error) by using temporal feature windows (feeding the MLPs with the features belonging to the past, current and future frames).

In 2008, Petridis and Pantic started to enrich the so far mainly audio-based work in laughter detection by consulting audio-visual cues for decision level fusion approaches [4]–[6]. They combined spectral and prosodic features from the audio modality with head movement and facial expressions from the video channel. Results suggest that integrated information from audio and video leads to improved classification reliability compared to a single modality - even with fairly simple fusion methods. They reported a classification accuracy of 74.7% to distinguish three classes, namely unvoiced laughter, voiced laughter and speech. In [7] they present a new classification approach for discriminating laughter from speech by modelling the relationship between acoustic and visual features with Neural Networks.

B. Acoustic laughter synthesis

Acoustic laughter synthesis is an almost unexplored domain. Only 2 attempts have been reported in literature. Sundaram and Narayanan [8] modeled the laughter intensity rhythmic envelope with the equations governing an oscillating mass-spring and synthesized laughter vowels by Linear Prediction. This approach to laughter synthesis was interesting, but the produced laughs were judged as non-natural by listeners. Lasarcyk and Trouvain [9] compared laughs synthesized by an articulatory system (a 3D modeling of the vocal tract) and diphone concatenation. The articulatory system gave better results, but they were still evaluated as significantly less natural than human laughs. In 2010, Cox conducted an online evaluation study to measure to what extent (copy-)synthesized laughs were perceived as generated by a human or a computer [10]. Laughs synthesized by the 2 aforementioned groups were included in the study, as well as a burst-concatenation copy-synthesized laughter proposed by UMONS, which obtained the best results with almost 60% of the 6000 participants thinking it could be a human laugh. Nevertheless, this number is far from the 80% achieved by a true human laugh.

C. Visual laughter synthesis

The audio-synchronous visual synthesis of laughter requires the development of innovative hybrid approaches that combine several existing animation techniques such as data-driven animation, procedural animation and machine learning based animation. Some preliminary audio-driven models of laughter have been proposed. In particular Di Lorenzo et al. [11] proposed an anatomic model of torso respiration during laughter, while Cosker and Edge [12] worked on facial animation during laughter. The first model does not work in real-time while the second is limited to only facial animation.

D. Laughing virtual agents

Urbain et al. [13] have proposed the AVLaughterCycle machine, a system able to detect and respond to human laughs in real time. With the aim of creating an engaging interaction loop between a human and the agent they built a system capable of recording the user's laugh and responding to it with a similar laugh. The virtual agent response is automatically chosen from an audio-visual laughter database by analyzing acoustic similarities with the input laughter. This database is composed of audio samples accompanied by the motion capture data of facial expressions. While the audio content is directly replayed, the corresponding motion capture data are retargeted to the virtual model.

Shahid et al. [14] proposed Adaptive Affective Mirror, a tool that is able to detect user's laughs and to present audio-visual affective feedback, which may elicit more positive emotions in the user. In more details, Adaptive Affective Mirror produces a distortion of the audio-visual input using real-time graphical filters such as bump distortion. These distortions are driven by the amount and type of user's laughter that has been detected. Fukushima et al. [15] built a system able to increase users' laughter reactions. It is composed of a set of toy robots that shake heads and play preregistered laughter sounds when the system detects the initial user laughter. The evaluation study showed that the system enhances the users' laughing activity (*i.e.*, generates the effect of contagion).

Finally, Becker-Asano et al. [16] studied the impact of auditory and behavioral signals of laughter in different social robots. They discovered that the social effect of laughter depends on the situational context including the type of task executed by the robot, verbal and nonverbal behaviors (other than laughing) that accompany the laughing act [17]. They also

claim that inter-cultural differences exist in the perception of naturalness of laughing humanoids [16].

III. Scenarios and Stimulus Film

In our evaluation scenario the virtual agent and its laughter behavior were investigated. The experimental setup involved a participant watching a funny video with a virtual agent visually present on a separate screen. The expressive behavior of the virtual agent was varied among three conditions, systematically altering the degree of expressed appreciation of the clip (amusement) in verbal and non-verbal behavior, as well as different degrees of interaction with the participant's behavior. The three conditions are:

- "fixed speech": the agent is verbally expressing amusement at pre-defined times of the video
- "fixed laughter": the agent is expressing amusement through laughs at pre-defined times of the video
- "interactive laughter": the agent is expressing amusement through laughter, in reaction to both the stimulus video and the participant's behavior

Furthermore, participant related variables were assessed with self-report instruments and allowed for the investigation of the influence of mood and personality on the perception and evaluation of the virtual agent. This allowed for the control of systematic biases on the evaluation of the virtual agent, which are independent of its believability (*e.g.*, individuals with a fear of being laughed at perceive all laughter negatively). The impact of the agent was assessed by investigating the influence of the session on participant's mood, as well as by self-report questionnaires assessing the perception of the virtual agent and the participant's cognitions, beliefs and emotions.

The stimulus film consisted of five candid camera pranks with a total length of 8 minutes. The clips were chosen by one expert rater who screened a large amount of video clips (approximately 4 hours) and chose five representative, culturally unbiased pranks sections of approximately 1 to 2 minutes length. All pranks were soundless and consisted of incongruity-resolution humor.

IV. Data used for training

Several pieces of data have been used in the project, two existing databases and two datasets specifically recorded to develop Laugh Machine. These databases are briefly presented in this section.

A. The SEMAINE database

The SEMAINE database [18] was collected for the SEMAINE-project by Queen's University Belfast with technical support of the HCI^2 group of Imperial College London. The corpus includes recordings from users while holding conversations with an operator who adopts in sequence four roles designed to evoke emotional reactions. One of the roles (Poppy) being happy and outgoing often invokes natural and spontaneous laughter by the user. The corpus is freely available for research purpose and offers high-audio quality, as well as, frontal and profile video recordings. The latter is important as it allows incorporation of visual features, which is part of the future work of Laugh Machine.

Within Laugh Machine, the SEMAINE database has been used to design a framework for laughter recognition (see Section VI-B) and select the most relevant audio features for this task.

Even though laughter is included as a class in the transcriptions of the SEMAINE database, provided laughter annotation tracks turned out to be too coarse to be used in the Laugh Machine training process. Hence, 19 sessions (each about 4-7 minutes long), which were found to include a sufficient number of laughs, were selected and manually corrected.

B. The AVLaughterCycle database

Secondly, we used the AudioVisualLaughterCycle (AVLC) corpus [19] that contains about 1000 spontaneous audio-visual laughter episodes with no overlapping speech. The episodes were recorded with the participation of 24 subjects. Each subject was recorded watching a 10-minutes comedy video. Thus it is expected that the corpus contains mainly amusement laughter. Each episode was captured with one motion capture system (either Optitrack or Zigntrack) and synchronized with the corresponding audiovisual sample. The material was manually segmented into episodes containing just one laugh. The number of laughter episodes for a subject ranges from 4 to 82. The annotations also include phonetic transcriptions of the laughter episodes [20].

Within Laugh Machine, the AVLaughterCycle database has been used to design the output components (audiovisual laughter synthesis, see Section VIII).

C. Belfast interacting dyads

The first corpus recorded especially for Laugh Machine contains human-human interactions when watching the stimulus film (see Secton III). Two dyads (one female-female, one male-male) were asked to watch the film. The two participants were placed in two rooms; they watched the same film simultaneously on two separate LCD displays. They could also see the other participant's reaction as a small window with the other person view was placed on the top of the displayed content. The data contains the closeup view of each participant's face, 90 degree views (all at 50FPS) of the half of the body as well as audio tracks obtained from close-talk and far-field microphones for each participant, sampled at 48kHz and stored in PCM 24bits. Laughs have been segmented from the recorded signals.

This interaction data has been used to train the dialog manager component (see Section VII).

D. Augsburg scenario recordings

In order to tune the laughter detection (initially developed on the SEMAINE database) to the sensors actually used in Laugh Machine, a dedicated dataset has been recorded.

Since laughter includes respiratory, vocal, and facial and skeletomuscular elements [21], we can expect to capture signs of laughter if we install sensory to capture the user's voice,

facial expressions, and movements of the upper body. To have a minimum of sensors we decided to work with only two devices: the Microsoft Kinect and the Chest Band developed at the University College London (see Section VI-D). The latest version of the Microsoft Kinect SDK[1] not only offers full 3D body tracking, but also a real-time 3D mesh of facial features—tracking the head position, location of eyebrows, shape of the mouth, etc.

TABLE I
RECORDED SIGNALS.

Recording device	*Captured signal*	*Description*
Microsoft Kinect	Video	RGB, 30fps, 640x480
	Face points	
	Facial action units	
	Head pose	
	Skeleton joints	
	Audio	16 kHz, 16 bit, mono
Respiration Sensor	Thoracic circumference	120Hz, 8 bit

The recorded signals are summarized in Table I. Recordings took place at the University of Augsburg, using the Social Signal Interpretation (SSI, see Section VI-A) tool. During the sessions 10 German and 10 Arabic students were recorded while watching the stimulus film. By including participants with different cultural background it is our hope to improve the robustness of the final system. The recordings were then manually annotated at three levels: 1) beginning and ending of laughter in the audio track, 2) any non-laughter event in the audio track, such as speech and other noises, and 3) beginning and ending of smiles in the video track.

V. SYSTEM ARCHITECTURE

The general system architecture is displayed in Figure 1. We can distinguish 3 types of components: input components, decision components and output components. They are respectively explained in Sections VI, VII and VIII.

The input components are responsible for multimodal data acquisition and real-time laughter-related analysis. They include laughter detection from audiovisual features, body movements analysis (with laughter likelihood), respiration signal acquisition (also with laughter likelihood) and input laughter intensity estimation.

The decision components receive the information from the input components (*i.e.*, laughter likelihoods and intensity from multimodal features) as well as contextual information (*i.e.*, the funniness of the stimulus, see Section IX-C2, in green on Figure 1) and determines how the virtual agent should react. There are actually two decision components: the dialog manager, which decides if and how the agent should laugh at each time frame (typically 200ms), is followed by a block called "Laughter Planner", which decides whether or not the instruction to laugh should be forwarded to the synthesis components. In some cases, for example when there is an ongoing animation, it is indeed preferable not to transmit new synthesis instructions.

[1]http://www.microsoft.com/en-us/kinectforwindows/

The output components are responsible for the audiovisual laughter synthesis that is displayed when the decision components instruct to do so. In the current state of these components, it is not possible to interrupt a laughter animation (*e.g.*, to decide abruptly to stop laughing or on the other hand to laugh more intensely before the current output laughter is finished). This is the reason why the "Laughter Planner" module has been added. The Laugh Machine project includes one component for audio synthesis and 2 different animation Realizers, Greta and LivingActor (see Section VIII).

All the components have to work in real-time. Thus, the organization of the communication between different components is crucial in such project. For this purpose we use the SEMAINE[2] architecture which was originally aimed to build a Sensitive Listener Agent (SAL). The SEMAINE API is a distributed multi-platform component integration framework for real-time interactive systems. The architecture of SEMAINE API uses a message-oriented middleware (MoM) in order to integrate several components – where actual processing of the system is defined. Such components communicate via a set of topics. Here, a topic is a virtual channel where each and every published message, addressed to that topic, is delivered to its subscribed consumers. The communication passes via the message-oriented middleware ActiveMQ™ [22], which supports multiple operating systems and programming languages. For component integration, the SEMAINE API encapsulates the communication layer in terms of components that receive and send messages, and a system manager that verifies the overall system state, provides a centralized clock independent of the individual system clocks.

To integrate Laugh Machine components we used the same exchange messages server (*i.e.*, ActiveMQ) and the SEMAINE API. Each Laugh Machine component can read and write to some specific ActiveMQ topics. For this purpose we defined a hierarchy of message topics and for each topic the appropriate message format. Simple data (such as input data or clock signals) were coded in simple text messages in string/value tuples, so called *MapMessage*s, *e.g.* the message $AUDIO\_LAUGHTER\_DETECTION$ 1 is sent wherever laughter was detected from the audio channel. On the other hand more complex information such as the description of the behavior to be displayed was coded in standard XML languages such as the Behavior Markup Language[3] (BML).

It should be noted that, since the available data to train the decision components ((*i.e.*, the Belfast dyads data) did not contain Kinect nor respiration signals, the decision modules currently use only the acoustic laughter detection and acoustic laughter intensity. The other input components (in yellow on Figure 1) are nevertheless integrated in the system architecture and their data is recorded in order to train the decision modules with these additional signals in the future.

VI. INPUT COMPONENTS

To work properly, our system must be able to capture sufficient information about the user, coming from different

[2]http://www.semaine-project.eu/

[3]http://www.mindmakers.org/projects/bml-1-0/wiki/Wiki?version=10

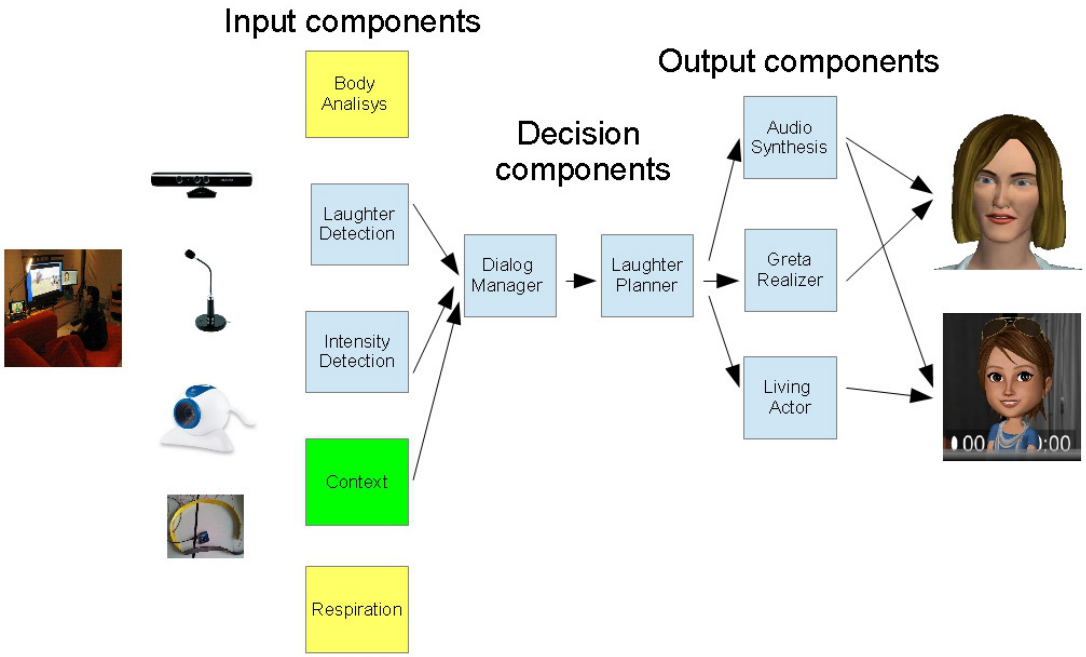

Fig. 1. Overall architecture of Laugh Machine

modalities such as sound, visual tracking, and chest movement. To facilitate the multimodal data processing and the synchronisation between the different signals, we have used the Social Signal Interpretation (SSI) [23] software developed at the University of Augsburg. This software will be presented first in this section, then we will present the different analysis components that have been developed: audiovisual laughter detection, laughter intensity estimation, respiration signal acquisition and body movement analysis. All these components have been plugged directly in SSI, except the body motion analysis, due to a problem of sharing the Kinect data in real-time.

A. SSI

The desired recognition component has to be equipped with certain sensory to capture multimodal signals. First, the raw sensor data is collected, synchronized and buffered for further processing. Then the individual streams are filtered, *e.g.* to remove noise, and transformed into a compact representation by extracting a set of feature values from the time- and frequency space. The in this way parameterized signal can be classified by either comparing it to some threshold or applying a more sophisticated classification scheme. The latter usually requires a training phase where the classifier is tuned using pre-annotated sample data. The collection of training data is thus another task of the recognition component. Often, an activity detection is required in the first place in order to identify interesting segments, which are subject to a deeper analysis. Finally, a meaningful interpretation of the detected events is only possible at the background of past events and events from other modalities. For instance, detecting several laughter events within a short time frame increases the probability that the user is in fact laughing. On the other hand, if we detect that the user is talking right now we would decrease the confidence for a detected smile. The different tasks the recognition component is involved with are visualized in Figure 2.

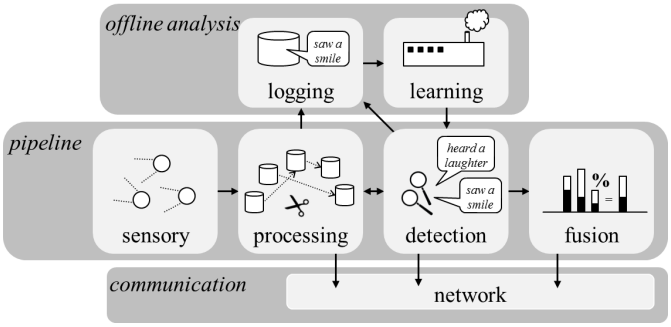

Fig. 2. Scheme of the laughter recognition component implemented with the Social Signal Interpretation (SSI) framework. Its central part consists of a recognition pipeline that processes the raw sensory input in real-time. If an interesting event is detected it is classified and fused with previous events and those of other modalities. The final decision can be shared through the network with external components. In order to train the recognition components a logging mechanism is incorporated in order to capture processed signals and add manual annotation. In an offline learning step the recognition components can now be tuned to improve accuracy.

The Social Signal Interpretation (SSI) software [23] developed at Augsburg University suits all mentioned tasks and was therefore used as a general framework to implement the recognition component. SSI provides wrappers for a large range of commercial sensors, such as web/dv cameras and multi-channel ASIO audio devices, as well as the Nintendo Wii remote control, Microsoft Kinect and various physiological sensors like NeXus, ProComp, AliveHeartMonitor, IOM or Emotiv. A patch-based architecture allows a developer to quickly construct pipelines to simultaneously manipulate the

raw signals captured by multiple devices, where the length of the processing window can be adjusted for each modality individually. Many common filter algorithms, such as moving and sliding average, Butterworth, Chebyshev, Elliptic, etc. as well as, derivative and integral filters are part of the core system and can be easily combined with a range of low-level features such as Fourier coefficients, intensity, cepstra, spectrogram, or pitch, as well as, more than 100 functionals, such as crossings, extremes, moments, regression, percentiles, etc. However, a plug-in system encourages developers to extend the core functions with whatever algorithm is required. A peak detection component is included, too, which can be applied to any continuous signal in order to detect segments above a certain activity level. If an event is detected it can be classified using one of various classification models such as K-Nearest Neighbor (KNN), Linear Discriminant Analysis (LDA), Support Vector Machines (SVM) or Hidden Markov Models (HMM). Tools for training and evaluation are available and can be combined with several feature selection algorithms (*e.g.*, SFS) and over-sampling techniques (*e.g.*, SMOTE [24]) for boosting under represented classes are available, too. Finally, classified events can be fused over time using vector-based event fusion. SSI offers a XML interface to put the different components to a single pipeline and keep control of important parameters.

In the Laugh Machine project, SSI was used for body and face tracking as well as audio and respiration recording. To have access to the new features provided in the latest Microsoft Kinect SDK, the Kinect wrapper in SSI was revised and updated accordingly. To access to the stretch values measured by the respiration sensor a new sensor wrapper was written using a serial connection.

After finishing the integration of the sensor devices, a recording pipeline was set up to record a training corpus for tuning the final recognition system (the Augsburg scenario recordings presented in Section IV-D). The pipeline also includes a playback component that allows replay of a video file to the user in order to induce laughter. This feature was used to drive the stimulus video directly from SSI. Since the video playback is then synchronized with the recorded signals, it is possible to relate captured laughter bouts to a certain stimuli in the video. The same pipeline was later used in our experiments. It is illustrated in Figure 3, which presents the Laugh Machine architecture from the point of view of SSI. The following sections present components that have been integrated into SSI: laughter detection, laughter intensity estimation and respiration signal acquisition.

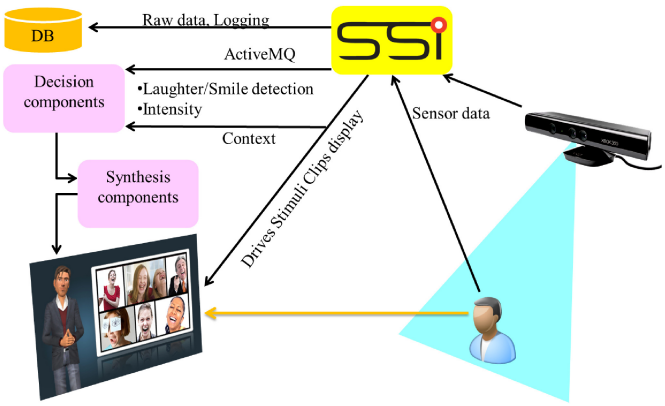

Fig. 3. SSI roles in the Laugh Machine system. While the user is watching funny video clips his or her non-verbal behavior is analyzed by a recognition component developed with SSI. If a laughter event is detected this information is shared to the behavior model, which controls the avatar engine. According to the input the avatar is now able to respond in an appropriate way, *e.g.*, join the user's laughter bout.

B. Laughter detection

Starting from the literature one can find several studies dealing with the detection of laughter from speech (*e.g.*, [1]–[3], see Section II-A). Most of them are pure offline studies and in part the explored feature types and classification methods vary largely. This circumstance makes it difficult to decide from scratch, which feature set and classifier would be the best choice for an online laughter detector. Hence, it was decided to run a fair comparison of the suggested methods in a large scale study. To this end, the SEMAINE database has been used and annotations of 19 files containing laughter were manually edited (as explained in Section IV-A).

Based on the edited annotations, for each second (a number commonly found in literature) it was decided whether the segment includes only silence (1906 samples), pure speech (5328), pure laughter (370), or both, speech and laughter (261). Samples were then equally distributed in a training and test set, while it was ensured that samples of the same user would not occur in both sets. To have an equal number of samples for each class, underrepresented classes were oversampled in the training set using SMOTE. After some preliminary tests it was decided to leave out silence, as it can be easily differed from speech and laughter using activity detection. It was also decided to leave out samples including both speech and laughter, as the goal of the experiment was to find features that best discriminate the two classes.

After setting up the database, large parts of the openSMILE (Speech & Music Interpretation by Large-space Extraction) feature extraction toolkit developed at the Technical University Munich (TUM) [25] were integrated into SSI. OpenSMILE is an open source state-of-the-art implementation of common audio features for speech and music processing. An important feature is its capability of on-line incremental processing, which makes it possible to run even complex and time-consuming algorithms, such as pitch extraction, in real-time. Based on the findings of earlier studies, the following speech-related low-level features were selected as most promising candidates: Intensity, MFCCs, Pitch, PLPs. On these the following 11 groups of functionals were tested: Zero-Crossings, DCT (Direct Cosine Transform) Coefficients, Segments, Times, Extremes, Means, Onsets, Peaks, Percentiles, Linear and Quadratic Regression, and Moments. Regarding classification, four well known methods were chosen: Naive Bayes (NB), Gaussian Mixture Models (GMM), Hidden Markov Models (HMM) and Support Vector Machines (SVM). Finally, the frame size at which low-level features are extracted was also altered.

A large scale experiment was then conducted. First, each of the 11 groups of functionals was tested independently with each of the four low-level feature types. In case of MFFCs also the number of coefficients was altered and higher-order derivatives (up to 4) were added. Results suggest that most reliable results are achieved using Intensity and MFCCs, while adding Pitch and PLP features did not improve results on the studied corpus. Among the functionals, Regression, Moments, Peaks, Crossings, Means and Segments are considered to carry most distinctive information. Regarding classification, SVM with a linear kernel clearly outperformed all other tested recognition methods. In terms of operation size accuracy was highest at a frame rate of 10ms with 2/3 of overlap. In the best case an overall accuracy of 88.2% at an unweighted average recall of 91.2% was obtained.

The developed laughter detection framework was then tuned to the specific Laugh Machine scenario and input components (*i.e.*, the audio is recorded by the Kinect), thanks to the Augsburg scenario recordings (see Section IV-D). The annotations of the audio tracks were used to re-train the laughter detector described above, with the features extracted in the Laugh Machine scenario conditions. The obtained laughter model was finally combined with a silent detection to filter out silent frames in the first place and classifying all remaining frames into laughter or noise. The frame size was set to 1 second with an overlap of 0.8 second, *i.e.* a new classification is received every 0.2 second. The annotations of the video tracks are meant for training an additional smile detector in the future. Same counts for the respiration signal (see Section VI-D), which in future will serve as a third input channel to the laughter detector.

C. Laughter intensity

Knowing the intensity of incoming laughs is important information to determine the appropriate behavior of the virtual agent.

In [26], naive participants have been asked to rate the intensity of laughs from the AVLaughterCycle database [19] on a scale from 1 (very low intensity) to 5 (very high intensity). One global intensity value had to be assigned to each laugh. Audiovisual features that correlate with these perceived global intensity have then be investigated.

Here, we wanted not only to estimate the global laughter intensity, after the laugh has finished, but to measure in real-time the instantaneous intensity. As a first step, only the audio modality was included. 49 acoustic laughs, produced by 3 subjects of the AVLaughterCycle database and distributed over the ranges of annotated global intensity values, have been continuously annotated in intensity by one labeler. Acoustic features have been extracted with the objective to predict the continuous intensity curves.

Figure 4 displays the manual intensity curve for one laugh, together with the automatic intensity prediction obtained from two acoustic features: loudness and pitch. The intensity curve is obtained by a linear combination between the maximum pitch and the maximum loudness values over a sliding 200ms window, followed by median filtering to smooth the curve. The overall trend is followed, even though there are differences, mostly at the edge of the manually spotted bursts, and the manual curve is smoother than the automatic one. Furthermore, the overall laughter intensity can be extracted from the continuous annotation curve: correlation coefficients between the median intensity scored by users and the intensity predicted from acoustic features are over 0.7 for 21 out of 23 subjects[4].

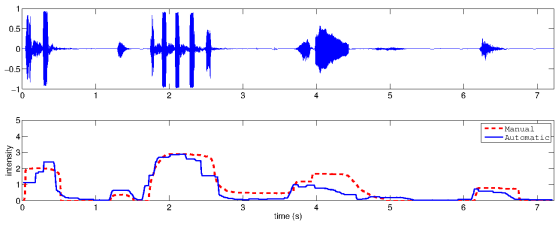

Fig. 4. Example of laughter continuous intensity curve. Top: waveform; Bottom: manual and automatic intensity curves.)

During the eNTERFACE workshop, work has been done to improve the computation of the continuous intensity curve. Indeed, the linear combination is able to capture trends for one subject (which laugh or laugh segment is more intense than another one), but the outputted values fall in different ranges from one subject to another. Classification with Weka [27] has been investigated to overcome this problem. First, neural networks have been trained in Weka to predict the continuous intensity curve from acoustic features (MFCCs and spectral flatness). The correlation with the manually annotated curves was over 0.8, using a "leave-one-subject-out" scheme for testing. Second, other neural networks have been used to compute the global laughter intensity from the predicted continuous intensity. To keep the number of features constants, 5 functionals (max, std, range, mean, sum) of the continuous intensity have been used as inputs. The results again show a good correlation between the predicted global intensity and the one rated by naive participants, in this case with similar values for all the subjects of the AVLaughterCycle database.

However, the speaker-independent intensity detection with Weka could not be integrated in the full LaughMachine system yet. Only the linear combination has been used in our experiments. Further work to improve laughter intensity prediction include the extension of the feature set to visual features, the integration of the Weka classification within the Laugh Machine framework and possibly the adaptation of the functions to the user.

D. Respiration

The production of audible laughter is, in essence, a respiratory act since it requires the exhalation of air to produce distinctive laughter sounds ("Ha") or less obvious sigh- or hiss-like verbalizations. The respiratory patterns of laughter have been extensively researched as Ruch & Ekman [21] summarize. A distinctive respiration pattern has emerged of

[4]The 24th subject of the AVLC corpus only laughed 4 times and all these laugsh were rated witht he same global intensity, which prevents us from computing correlations for this subject

a rapid exhalation followed by a period of smaller exhalations at close-to-minimum lung volume. This pattern is reflected by changes in the volume of the thoracic and abdominal cavities, which rapidly decrease to reach a minimum value within approximately 1 s [28]. These volumetric changes can be seen through the simpler measure of thoracic circumference, noted almost a century ago by Feleky [29]. In order to capture these changes, we constructed a respiration sensor based on the design of commercially available sensors: the active component is a length of extensible conductive fabric within an otherwise inextensible band that is fitted around the upper thorax. Expansions and contraction of the thorax change the length of the conductive fabric causing changes in its resistance. These changes in resistance are used to modulate an output voltage that is monitored by the Arduino prototyping platform[5]. Custom-written code on the Arduino converts the voltage to a 1-byte serial signal, linear with respect to actual circumference, which is passed to a PC over a USB connection at a rate of approximately 120Hz.

Automatic detection of laughter from respiratory actions has previously been investigated using electromyography (EMG). Fukushima et al. analyzed the frequency characteristics of diaphragmatic muscle activity to distinguish laughter, which contained a large high-frequency component, from rest, which contained mostly low-frequency components [15]. We exploited the predictable respiration pattern of laughter to use simpler techniques that do not rely on computationally demanding frequency decomposition. We identified laughter onset through the appearance of 3 respiration events (see Figure 5):

1) A sharp change in current respiration state (inhalation, pause, standard exhalation) to rapid exhalation.
2) A period of rapid exhalation resulting in rapid decrease in lung volume.
3) A period of very low lung volume

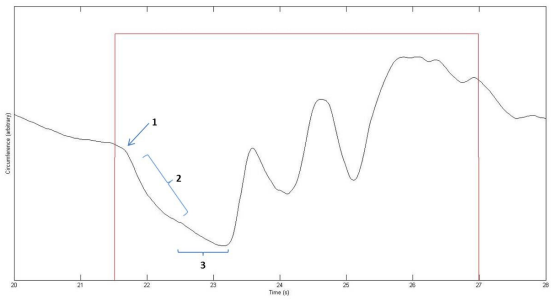

Fig. 5. Example of thoracic circumference, with laughter episode marked in red, and notable features of laughter initiation. Feature 1 - a sharp change in current respiration state to rapid exhalation; feature 2 - a period of rapid exhalation; feature 3 - a period of very low lung volume.

These appear as distinctive events in the thoracic circumference measure and its derivatives:

1) A negative spike in the second derivative of thoracic circumference.
2) A negative period in the first derivative of thoracic circumference.
3) A period of very low thoracic circumference.

These were identified by calculating a running mean (λ_f) and standard deviation (σ_f) for each measure. A running threshold (T_f) for each measure was calculated as:

$$T_f = \lambda_f - \alpha_f \sigma_f \quad (1)$$

where α_f is a coefficient for that measure, empirically determined to optimise the sensitivity/specificity trade-off. Each feature was determined to be present if the value of the measure fell below the threshold at that sample. Laughter onset was identified by the presence of all three features in the relevant order (1 before 2 before 3) in a sliding window of approximately 1 s. This approach restricts the number of parameters to 3 ($\alpha_{1\text{—}3}$) but does introduce lag necessary for calculating valid derivatives from potentially noisy data. It also requires a period for the running means and standard deviations, and so the running thresholds, to stabilise. This process would be jeopardised by the presence of large, rapid respiratory event such as coughs and sneezes. We were unable to integrate these rules into the LaughMachine system due to technical errors. Future recordings on the LaughMachine platform, incorporating the respiration data, will allow optimisation of these rules and the fusion of respiration data with other modalities for real-time laughter/non-laughter discrimination.

E. Body analysis

The EyesWeb XMI platform is a modular system that allows both expert (e.g., researchers in computer engineering) and non-expert users (e.g., artists) to create multimodal installations in a visual way [30]. The platform provides modules, called blocks, that can be assembled intuitively (*i.e.*, by operating only with mouse) to create programs, called patches, that exploit system's resources such as multimodal files, webcams, sound cards, multiple displays and so on. The body analysis input component consists of an EyesWeb XMI patch performing analysis of the user's body movements in realtime. The computation performed by the patch can be split into a sequence of distinct steps, described in the following subsections.

1) Shoulder tracking: The task of the body analysis module is to track the user's shoulders and perform some computation on the variation of their position in realtime. In order to do that we could provide the Kinect shoulders' data extracted by SSI (see Section VI-B) as input to our component. However, we observed that the shoulders' position extracted by Kinect do not consistently follow the user's real shoulder movement: in the Kinect skeleton, shoulders' position is determined by performing some statistical algorithm on the user's silhouette and depth map and usually this computation can not track subtle shoulders' movement, for example, small upward/downward movements. To overcome this limitation we fixed two markers on the user's body: two small and lightweight green polystyrene spheres have been fixed on the user's clothes just over the user's shoulders. The EyesWeb patch separates the green channel of the input video signal

[5]http://www.arduino.cc/

to isolate the position on the video frame of the two spheres. Then a tracking algorithm is performed to follow the motion of the sphere frame by frame, as shown in Figure 6.

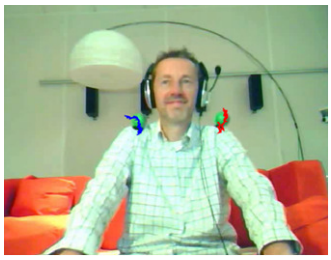

Fig. 6. Two green spheres placed on the user's shoulders are tracked in realtime (red and blue trajectories)

The position of each user's shoulder is associated to the barycenter of each sphere, which can be computed in two ways. The first consists in the computation of the graphical barycenter of each sphere, that is, the mean of the pixels of each sphere's silhouette is computed. The second option includes some additional steps: after computing the barycenter like in the first case, we consider a square region around it and we apply a Lukas-Kanade [31] algorithm to this area. The result is a set of 3 points on which we compute the mean: the resulting point is taken as the position of the shoulder.

2) Correlation: Correlation ρ is computed as the Pearson correlation coefficient between the vertical position of the user's left shoulder and the vertical position of the user's right shoulder. Vertical positions are approximated by the y coordinate of each shoulder's barycenter extracted as mentioned above.

3) Kinetic energy: It is computed from the speed of user's shoulders and their percentage mass as referred by [32] :

$E = \frac{1}{2}(m_1 v_1 + m_2 v_2)$

4) Periodicity: Kinetic energy is serialized in a sliding window time-series having a fixed length. Periodicity is then computed on such time-series, using Periodicity Transforms [33]. The input data is decomposed into a sum of its periodic components by projecting data onto periodic subspaces. Periodicity Transforms also provide a measure of the relative contribution of each periodic signal to the original one. Among many algorithms for computing Periodicity Transforms, we chose mbest. It determines the m periodic components that, subtracted from the original signal, minimize residual energy. With respect to the other algorithms, it also provides a better accuracy and does not need the definition of a threshold.

5) Body Laughter Index: Body Laughter Index (BLI) stems from the combination of the averages of shoulders' correlation and kinetic energy, integrated with the Periodicity Index. Such averages are computed over a fixed range of frames. However such a range could be automatically determined by applying a motion segmentation algorithm on the video source. A weighted sum of the mean correlation of shoulders' movement and of the mean kinetic energy is carried out as follows:

$BLI = \alpha\bar{\rho} + \beta\bar{E}$

As reported in [21], rhythmical patterns produced during laughter usually have frequencies around 5 Hz. In order to take into account such rhythmical patterns, the Periodicity Index is used. In particular, the computed BLI value is acknowledged only if the mean Periodicity Index belongs to the arbitrary range $[\frac{fps}{8}, \frac{fps}{2}]$, where fps is the input video frame rate (number of frames per second).

6) ActiveMQ: The EyesWeb XMI platform can be expanded to implement new functionalities that could be included into new sets of programming modules (blocks). To allow the communication between the body analysis patch and the other components (e.g., the SSI audio and face analysis component) we implemented two new blocks: the ActiveMQ receiver and the ActiveMQ sender. Body analysis component sends two types of data using the ActiveMQ message format described in Section V: data messages and clock messages. Data messages contain tuples representing the values of the user's shoulders movement features presented above. Clock messages contain the system clock of the machine on which the EyesWeb XMI platform is running. They are sent to the ActiveMQ server on which all the other components are registered. So, the local clock of all the components (audio and face analysis, dialogue generation and so on) is constantly updated with the same value and synchronization between the different component can be assured. In the future we aim to exploit the synchronization features embedded in the SEMAINE platform, that is implemented as a layer of the ActiveMQ communication protocol.

VII. Dialog manager

The laughter-enabled dialogue management module aims at deciding, given the information from the input components (*i.e.*, laughter likelihoods and intensity from multimodal features) as well as contextual information (*i.e.*, the funniness of the stimulus), when and how to laugh so as to generate a natural interaction with human users. In this purpose, the dialogue management task is seen as a sequential decision making process meaning that the behavior is not only influenced by the current context but also by the history of the dialogue. This is a main difference comparing with the other interactive systems such as SEMAINE. The optimal sequence of decisions is learned from actual human-human or human-computer interaction data and is not rule-based or handcrafted which is another difference with the SEMAINE system.

The decision of whether and how to laugh must be taken at each time frame. For the eNTERFACE workshop a time frame lasts $\Delta t = 200$ms. The input I received by the Dialog manager at each time frame is a vector ($I \in [0,1]^k$: each feature has been normalized) where k is the number of chosen multimodal features. The output O produced at each time frame is a vector ($O \in [0,1] \times [0, \text{time}_{\text{max}}]$) where the first dimension codes the laughter intensity and the second dimension codes the duration of the laugh.

The method used to build the decision rule during the eNTERFACE workshop is a supervised learning method. A supervised learning method is able via a training data set $D = \{x_i, y_i\}_{1 \leq i \leq J}$ ($\{x_i\}_{1 \leq i \leq J}$ are the inputs which belong to the set X, $\{y_i\}_{1 \leq i \leq J}$ are the labels which belong to the set Y and $J \in \mathbb{N}^*$) to build a decision rule π. The decision rule π is a function from X to Y that generalizes the relation between

the inputs x_i and the labels y_i of the training data set. There are two different types of supervised methods: Classification when the number of outputs is finite and Regression when the number of outputs is infinite.

A. Training of the Dialog manager

To apply a supervised learning method to our dialog manager, we need a training data set specific to our scenario (see Section III). The Belfast interaction dyads (see Section IV-C) was recorded to this purpose. Let us name the two interacting participants $P1$ and $P2$, respectively recorded on tracks $T1$ and $T2$. We recall that the participants watch simultaneously the same stiumulus video, and can also see (and hear) each other on the display screen: $P2$ is viewable by $P1$ and is considered as playing the role of the virtual agent. The length of a recording is $H = K\Delta t$. Thus, on $T1$ we have the inputs (*i.e.*, laughter likelihoods and intensity from multimodal features of $P1$) $\{I_i\}_{1\leq i\leq K}$ of the Dialog manager and on $T2$ the corresponding outputs $\{O_i\}_{1\leq i\leq K}$ which are the intensities and durations of the laughs of $P2$.

The aim of the supervised method is to find a decision rule such that the virtual agent will be able to imitate $P2$. Before applying a supervised method, we decided to cluster the inputs with N clusters (via a k-means method) and to cluster the outputs with M clusters (via a Gaussian Mixture Model, or GMM, method). k-means clustering is a method of cluster analysis which aims to partition $n \in \mathbb{N}^*$ observations into $0 \leq k \leq n$ clusters in which each observation belongs to the cluster with the nearest mean. This results in a partitioning of the data space into Voronoi cells. GMM clustering is a method of cluster analysis where each cluster can be parameterized by a Gaussian distribution. The choice of the GMM method for the output clustering is explained in Section VII-B.

Thanks to clustering, the input data becomes the input clustered data $\{I_i^C\}_{1\leq i\leq K}$ with $I_i^C \in \{1,\dots,N\}$ and the output data becomes the output clustered data $\{O_i^C\}_{1\leq i\leq K}$ with $O_i^C \in \{1,\dots,M\}$. Clustering the inputs allows to have a finite decision rule which means that the decision rule can be represented by a finite vector. Clustering the outputs allows using a classification method such as the k-nearest neighbors (k-nn) instead of a regression method which is more difficult to implement.

Finally the supervised method used on the clustered data $\{I_i^C, O_i^C\}_{1\leq i\leq K}$ is a k-nearest-neighbor method which gives us the decision rule π which is a function from $\{1,\dots,N\}$ to $\{1,\dots,M\}$. k-nn is a method for classifying objects based on closest training examples: the object is assigned to the most common label amongst its k nearest neighbors. Figure 7 represents the training phase of the Dialog manager needed to obtain the decision rule π.

B. Using the Dialog manager

The decision rule π obtained by the classification method on $\{I_i^C, O_i^C\}_{1\leq i\leq K}$ is a function from $\{1,\dots,N\}$ to $\{1,\dots,M\}$: it takes an input cluster and it gives an output cluster. However, our dialog manager must be able to take an input $I \in [0,1]^k$ and give an output $O \in [0,1] \times [0, \text{time}_{\max}]$.

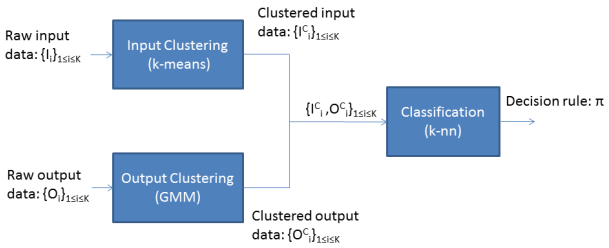

Fig. 7. Dialog Manager training

So, first we need to assign the input $I \in [0,1]^k$ to the corresponding input cluster $I^C \in \{1,\dots,N\}$. To do that, we choose the cluster for which the mean is the closest to $I \in [0,1]^k$:

$$I_C = \underset{1\leq i\leq N}{\operatorname{argmin}} \|I - \mu_i^I\|_2^2, \tag{2}$$

where μ_i^I is the mean of the input cluster $i \in \{1,\dots,N\}$ and $\|\|_2$ is the euclidean norm. This operation is called the input cluster choice. Second, to be able to generate O from the selected output cluster $l \in \{1,\dots,M\}$ the question is: which element of the output cluster l must we choose in order to correspond to the data $\{O_i\}_{1\leq i\leq K}$? This is why we use a GMM method for clustering the outputs: each cluster l can be seen, in the 2-dimensional intensity-duration plane, as a Gaussian of law $\mathcal{N}(\mu_l^O, \Sigma_l^O)$, where μ_l^O is the mean of the output cluster l and Σ_l^O is the covariance matrix of the output cluster l. Therefore, to obtain an output, it is sufficient to sample an element O of law $\mathcal{N}(\mu_l^O, \Sigma_l^O)$. This operation is called the output generation.

Let us summarize the functioning of the Dialog manager (see also Figure 8): we receive the input I, we associate this input to its corresponding input cluster $I^C \in \{1,\dots,N\}$, then the decision rule π gives the output cluster $\pi(I^C) \in \{1,\dots,M\}$, finally the output O is chosen in the output cluster $\pi(I^C) \in \{1,\dots,M\}$ via the output generation.

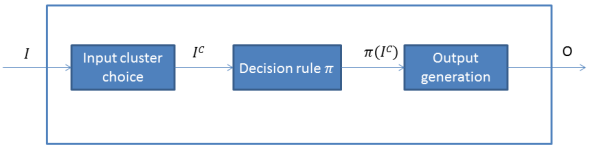

Fig. 8. Dialog Manager functioning

C. Laughter Planner

In the Laugh Machine architecture, the dialog manager is followed by the Laughter Planner, which is adapting the outputs of the dialog manager to the constraints (instruction format, avoid conflicting information, etc.) of the synthesis modules. While it technically is a decision component, the explanations about the Laughter Planner are included in the visual synthesis section (Section VIII-B).

VIII. Audiovisual Laughter synthesis

A. Acoustic laughter synthesis

Given 1) the lack of naturalness resulting from previous attempts to laughter acoustic synthesis, 2) the need for high level control of the laugh synthesizer and 3) the good performance achieved with Hidden Markov Model (HMM) based speech synthesis [34], we decided to investigate the potential of this technique for acoustic laughter synthesis. We opted for the HMM-based Speech Synthesis System (HTS) [35], as it is free and widely used in speech synthesis and research.

Explaining the details of speech synthesis with HMMs or HTS going beyond the scope of this project report, we will here only describe the major modifications that have been brought to adapt our laughter data to HTS and vice-versa, adapting functions or parameters of the HTS demo (provided with the HTS toolbox) to improve the quality of laughter synthesis. Readers who would like to know more about HTS are encouraged to consult the publications listed on the HTS webpage (http://hts.sp.nitech.ac.jp/?Publications), and in particular [34] for an overview or Yoshimura's Phd Thesis [36] for more detailed explanations.

The following paragraphs respectively focus on the selection of the training data, the modifications implemented in the HTS demo and, finally, the resulting process for acoustic laughter synthesis.

1) Selection and adaptation of acoustic data:

HMM-based acoustic synthesis requires a large quantity of data: statistical models for each unit (in speech: phonemes) can only be accurately estimated if there are numerous training examples. Furthermore, the data should be labeled (*i.e.*, a phonetic transcription must be provided) and come from a single person, whose voice is going to be modeled. HMM-based speech synthesis is usually trained with hours of speech.

It is difficult to obtain such large quantities of spontaneous laughter data. The only laughter database including phonetic transcriptions is the AVLaughterCycle database [19], [20], which contains in total 1 hour of laughter from 24 subjects. We decided to use that database for our acoustic laughter synthesis.

To fully exploit the potential of HTS, the phonetic annotations of the AVLaughterCycle database have been extended to syllables. Indeed, HTS is able to distinguish contexts that lead to different acoustic realizations of a single phoneme (and on the other hand, HTS groups the contexts that yield acoustically similar realizations of a phoneme). In speech, the context of a phoneme is defined not only with the surrounding phonemes, but also with prosodic information such as the position of the phoneme within the syllable, the number of phonemes in the previous, current and following syllables; the number of syllables in the previous, current and following words; the number of words in the phrase; etc. Except from the surrounding phones[6], such contextual information was not available in the AVLC annotations, as there was no annotation of the laughs in terms of syllables or words. It was decided to add a syllabic annotation of the data to provide the biggest possible contextual information. There is no clear definition of laughter syllables, and the practical definition that has been used for the syllabic annotation was to consider one syllable as a set of phones that was acoustically perceived as forming one block (or burst), usually containing one single vowel (but not always, as laughter can take different structures from speech). Since the syllabic annotation is time-consuming, it was decided to do it only for the subjects who laughed the most in the AVLaughterCycle database: subjects 5, 6, 14, 18 and 20. These subjects laugh around 5 minutes each, which is already far from the hours of training data used in speech synthesis, and it seemed they represent the best hopes for good quality laughter synthesis. The HTS contextual information was then formed by assimilating a full laughter episode to a speech sentence and laughter exhalation and inhalation segments to words.

In addition, due to the limited available data, the phonetic labels have been grouped in 8 broad phonetic classes—namely: fricatives, plosives, vowels, hum-like (including nasal consonants), glottal stops, nareal fricatives (noisy respiration airflow going through the nasal cavities), cackles (very short vowel similar to hiccup sound) and silence—instead of the 200 phones annotated in the AVLaughterCycle database [20]. Indeed, most of these phones had very few examples for each speaker, and hence could not be accurately modeled. Grouping acoustically similar phones enables to obtain better models, at the cost of reduced acoustic variability (*e.g.*, all the vowels are grouped in an average model that is close to 'a', and we loose the possibility to generate the few 'o's in the database).

An example of the resulting phonetic transcription is presented in Figure 9.

Finally, the laughs from the AVLaughterCycle database have been processed to reduce background noise and remove saturations.

2) Modifications of the HTS demo process:

Several minor modifications have been applied to HTS. Some of them are simple parameter variations compared to the standard values used in speech (and in the HTS demo). For example, the boundaries for fundamental frequency estimation have been extended (the values have been manually determined for each subject), the threshold for pruning decision trees has been increased, etc. In addition the list of questions available to decision trees has been extended, considering the new contextual information available for laughter.

More important, two standard HTS algorithms have been replaced by more efficient methods. First, the standard Dirac pulse train for voiced excitation has been replaced by the DSM model [37], which better fits the human vocal excitation shapes and reduces the buzziness of the synthesized voice. Second, the standard vocal tract and fundamental frequency estimation algorithms provided by HTS have been replaced by the STRAIGHT method [38], which is known in speech processing to provide better estimations.

3) Synthesis process:

With the explained modifications to the AVLaughterCycle database and the HTS demo, we were able to train laughter synthesis models, with which we can produce acoustic laughs

[6]Since the phonological notion of "phoneme" is not clearly defined for laughter; we prefer to use the word "phone" for the acoustic units found in our laughter database.

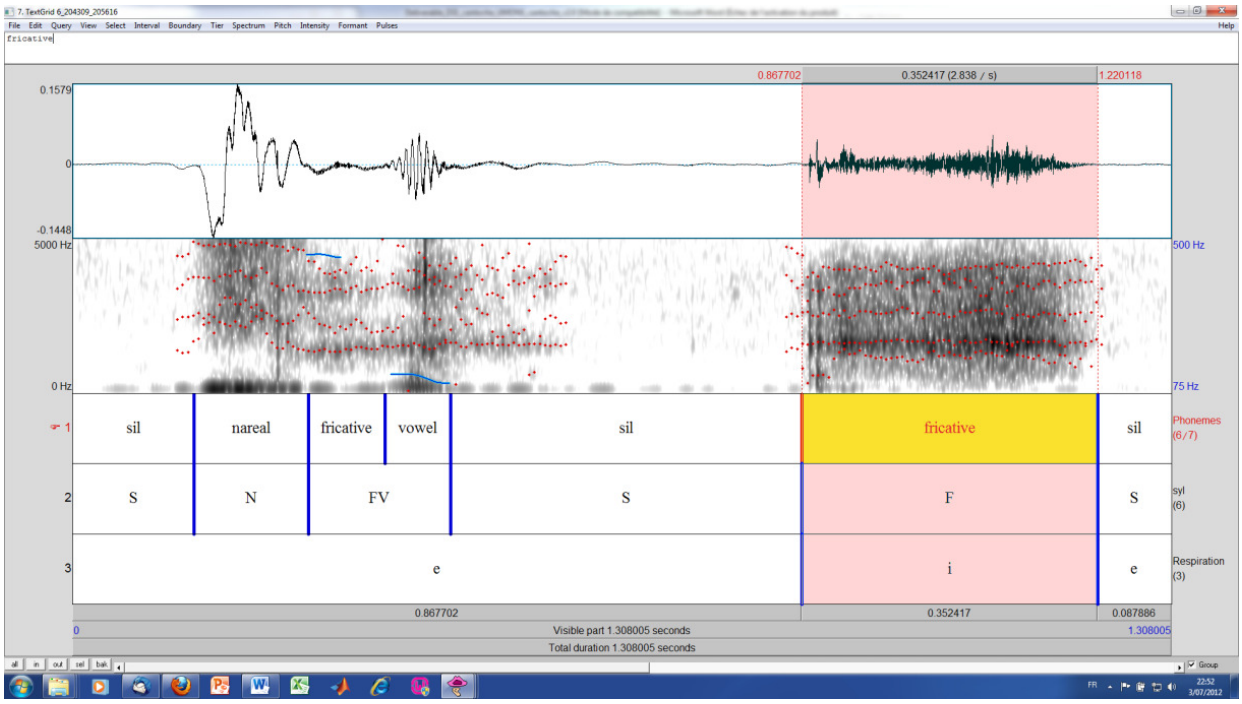

Fig. 9. Laughter phonetic and syllabic annotation: from top to bottom: a) waveform b) spectrogram c) phonetic annotation (using the 8 broad classes) d) syllable annotation e) respiration phases (inhalation or exhalation)

when giving an acoustic laughter transcription as input. It is worth noting that there is currently no module to generate such laughter phonetic transcriptions from high-level instructions (*e.g.*, a type of laughter, its duration and its intensity). We are thus constrained to play existing laughter transcriptions. Additionally, we noticed that the synthesis quality drops if we want to synthesize a phonetic transcription from speaker A with the models trained on voice B. In consequence, we currently stick to re-synthesizing laughs from one speaker, using both the phonetic transcription and the models trained from the same subject.

A perceptive evaluation study still has to be carried out. Nevertheless, the first qualitative tests are promising. The modifications explained in the previous paragraphs largely improved the quality of the laughter synthesis. There remain some laughs or sounds that are not properly synthesized, possibly due to the limited training data. Future works will investigate this issue as well as the possibility to generate new laughter phonetic transcriptions (or modify existing ones) that can be synthesized properly. Nevertheless, at the end of this project, we are able to synthesize a decent number of good quality laughs for the best voices coming from the AVLaughterCycle database.

B. Visual laughter Synthesis

Two different virtual agents and four different approaches were used for the visual synthesis. The visual synthesis component is composed of a Laughter Planner and 2 Realizers and Players (see Figure 10).

The Laughter Planner receives from the dialog manager the information about the appropriate laugh reaction through the ActiveMQ/SEMAINE architecture (see Section VII). Next it chooses one laugh episode from the library of predefined laugh samples and generates the appropriate BML command that is sent through ActiveMQ/SEMAINE to one out of two realizers available in the project: Living Actor or Greta Realizer.

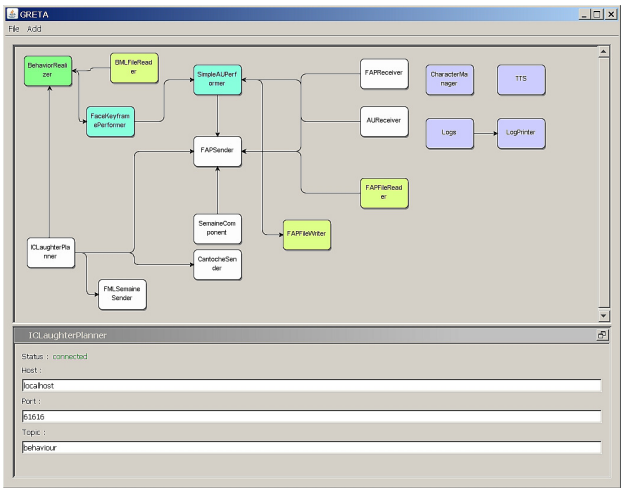

Fig. 10. Visual Synthesis Component Pipeline

On Figure 10 we present the detailed processing pipeline of our visual synthesis component. The Laughter Planner is connected to the Greta Behavior Realizer and the Cantoche Sender. The latter is responsible for the communication with the Living Actor component (see Section VIII-B4). Both Behavior Realizer and Cantoche Sender receive the same BML message. As these realizers use completely different methods for controlling the animation (Greta can be controlled by high-

level facial behavior description in FACS and low-level facial animation parameterization (MPEG-4/FAPs) while Living Actor plays predefined animations) we use realizer-specific extensions of BML to assure that the animations played with different agents are similar. If necessary, the Laughter Planner can also send commands in a high-level language called FML (FMLSemaineSender box) or control facial animations at very low level by specifying the values of facial animation parameters (FAPs) (FAPSender box). Independently of which of these pipelines is used the final animation is described using low level facial animation parameters (FAPs) and is sent through ActiveMQ/SEMAINE to the visualization module (FAPsender box). At the moment we use the Player from the SEMAINE Project. Four characters are included in this Player (2 male, 2 females) but for the purpose of the evaluation we used only one of them.

The Laughter Planner module can work in three different conditions, related to the three experimental scenarios: fixed speech condition (FSC), fixed laughter condition (FLC) and interactive laughter condition (ILC). In the first two conditions (FSC and FLC), the Laughter Planner receives the information about the context (time of funny event, see Section IX-C2) and it sends the agent verbal (FSC) or nonverbal (FLC) reaction pre-scripted in BML to be displayed to the user. The list of these behaviors was chosen manually.

In ILC condition the behavior of the agent is flexible as it is adapted to the participant and the context. The Laughter Planner receives the information on duration and intensity of laughter responses and using these values it chooses one laugh episode from the library that matches the best both values.

At the moment, the synthesis components do not allow for interruptions of the animation. Once it is chosen, the laugh episode has to be played until the end. During this period the Laughter Planner does not take into the account any new information coming from dialog manager. All the episodes start and end with a neutral expression. Thus they cannot be concatenated without passing through neutral face. Additionally the presynthesized audio wave file was synchronized with the animation.

Four different approaches were used in the project to prepare the lexicon of laughs: animation from the manual annotation of action units; animation from automatic facial movements detection; motion capture data driven; and manual animation. They are explained in the next subsections.

1) Animation from manual Action Units:

The Facial Action Coding System (FACS; [39]) is a comprehensive anatomically based system for measuring all visually discernible facial movement. It describes all distinguishable facial activity on the basis of 44 unique Action Units (AUs), as well as several categories for head and eye position movements and miscellaneous actions. Facial expressions of emotions are emotion events that comprise of a set of different AUs expressed simultaneously. Using FACS and viewing digital-recorded facial behavior at frame rate and in slow motion, certified FACS coders are able to distinguish and code all possible facial expressions. Utilizing this technique, a selection of twenty pre-recorded, laboratory stimulated, laughter events were coded. These codes were then used to model the facial behavior on the agent.

Four subjects interacting in same sex dyads watching the stimulus videos (see Section IV-C) were annotated by one certified FACS coder. Inter rater reliability was obtained by the additional coding of 50% of the videos by a second certified coder. The inter-rater reliability was sufficient (r = .80) and consent was obtained on events with disagreement. Furthermore, a selection of 20 laughter events from the AVLC laughter database [19] (subject 5) were coded by one certified coder.

The Greta agent is able to display any configuration of action units. For 3 characters (two females—Poppy and Prudence— and one male—Obadiah) single action units were defined and validated by certified FACS coders. A BML language implemented in Greta permits to control independently each action unit of the agent (its duration and intensity).

Furthermore, as a quality control, the animated AUs of the virtual agent was scrutinized by the FACS coders for a) anatomical appearance change accuracy, b) subtle differences and dominance rules relating to changes in the face when different intensity of facial expressions are produced.

During eNTERFACE we also developed a tool that automatically converts manual FACS annotation files to BML. Consequently any file containing manual annotation of action units can be easily displayed with the Greta agent.

2) Animation from Automatic Facial Movements detection: Greta uses Facial Animation Parameters (FAPs) to realize low level facial behavior. FAPs in Greta framework are represented as movements of MPEG-4 facial points compared to 'neutral' face. In order to estimate FAPs of natural facial expressions, we make use of an open-source face tracking tool—FaceTracker [40]—to track facial landmark localizations. It uses a Constrained Local Model (CLM) fitting approach that includes Regularized Landmark Mean-Shift (RLMS) optimization strategy. It can detect 66 facial landmark coordinates within real-time latency depending on system's configuration. Figure 11 shows an example of 2D and 3D landmark coordinates predicted by FaceTracker.

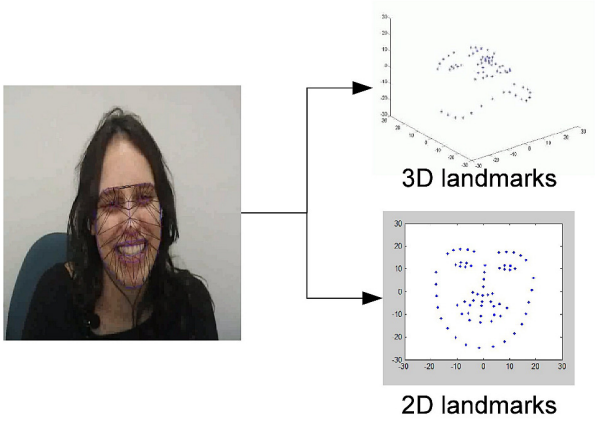

Fig. 11. Landmarks estimated by FaceTracker

Facial geometry is different for one and another. Therefore, it is difficult to estimate FAPs without neutral face calibration. To compute FAPs from facial landmarks, a neutral face model

is created with the help of 50 neutral faces of different persons. With the help of this model, FAPs are estimated as the distance between facial landmarks and neutral face landmarks. In case of user-specific FAP estimation in real-time scenario, the neutral face is estimated from a few seconds of video by explicitly requesting the user to be neutral. However, the better estimation of FAPs requires manual intervention for tweaking weights to map landmarks and FAPs, which is a down-side of this methodology. Figure 12 shows comparison of the locations between MPEG-4 FAP standard and the FaceTracker's landmark localizations.

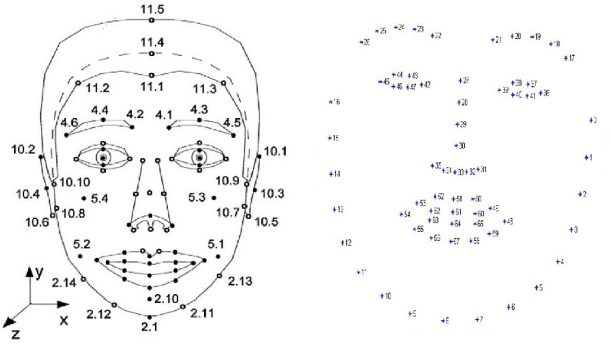

Fig. 12. (a) MPEG-4 FAP standard [left];(b) FaceTracker's landmark locations [right].

The landmark coordinates produced by the FaceTracker are observed as noisy due to the discontinuities and outliers in each facial point localization. Especially, the realized behavior is unnatural when we re-target the observed behavior onto Greta. In order to smooth the face tracking parameters, a temporal regression strategy is applied on individual landmarks by fitting 3rd order polynomial coefficients on a sliding window, where the sliding window size is 0.67 seconds (*i.e.*, 16 frames) and sliding rate is 0.17 seconds (*i.e.*, 4 frames). An example of the final animation can be seen on Figure 13.

3) Animation from Motion Capture Data:

The AVLC corpus (see Section IV-B) contains motion capture data of laugh episodes that has to be retargeted to the virtual model. The main problem in this kind of approaches consists in finding appropriate mappings for each participant's face geometry and different virtual models. Existing solutions are typically linear (*e.g.*, methods based on blendshape mapping) and do not take into account dynamical aspects of the facial motion itself. Recently Zeiler et al. [41] proposed to apply variants of Temporal Restricted Boltzmann Machines[7] (TRBM) to facial retargeting problem. TRBM are a family of models that permits tractable inference but allows complicated structures to be extracted from time series data. These models can encode a complex nonlinear mapping from the motion of one individual to another which captures facial geometry and dynamics of both source and target. In the original application [41] these models were trained on a dataset of facial motion capture data of two subjects, asked to perform a set of isolated facial movements based on FACS. The first subject had 313 markers (939 dimensions per frame) and the second subject had 332 markers (996 dimensions per frame). Interestingly there was no correspondence between marker sets.

We decided to use TRBM models for our project which involves retargeting from an individual to a virtual character. In our case, we take as input the AVLC mocap data and output the corresponding facial animation parameters (FAP) values. This task has two interesting aspects. First, the performance of these models was previously evaluated only on retargeting an isolated slow expression whereas our case involves transitions from laughing to some other expression (smile or idleness) as well as very fast movements. Second, we use less markers comparing to the original application. Our mocap data had only 27 markers on the face which is very sparse.

So far we used the AVLC data on one participant (number 5) as a source mocap data. We used two sequences, one of 250 frames and another one of 150 frames, to train this model. Target data (*i.e.*, facial animation parameters) for this training set was generated using the manual retargeting procedure explained in [13]. Both the input and output data vectors were reduced to 32 dimensions by retaining only their first 32 principal components. Since this model typically learns much better on scaled data (around [-1,1]), the data was then normalized to have zero mean and scaled by the average standard deviation of all the elements in the training set. Having trained the model, we used it to generate facial animation parameters values for 2 minutes long mocap data (2500 frames coming from the same participant). The first results are promising but more variability in the training set is needed to retarget more precisely different type of movements. It is important to notice that this procedure needs to be repeated for each virtual model (*e.g.*, Poppy, Prudence, Obadiah).

4) Manual Animation:

The Laugh Machine Living Actor module is composed of a real-time 3D rendering component using Living Actor technology and a communication component that constitutes the interface between the Living Actor agent and the ActiveMQ messaging system. Two virtual characters have been chosen for the first prototype: a girl and a boy, both with cartoonish style. Two types of laughter animations were created for each one by 3D computer graphics artists by visually matching the real person movies from the video database of interacting dyads (see Section IV-C).

Laughter capability has been added to the Living Actor character production tools and rendering component: specific facial morphing data are exported from 3D character animation tools and later rendered in real time. Laughter audio can be played from an audio file, which can either be the recording of a human laughter or a synthetic laughter synchronized with the real laughs. A user interface has been added to test various avatar behaviors and play sounds.

An Application Programming Interface has been added to the Laugh Machine Living Actor module to remotely control the avatar using BML scripts. A separate component was created in Java to make the interface between the Laugh Machine messaging system using ActiveMQ and TCP/IP messages of Living Actor API. At this stage, the supported BML instructions are restricted to a few commands, triggering

[7]The source code for these models is publicly available at http://www.matthewzeiler.com/software/RetargetToolbox/Documentation/index.html

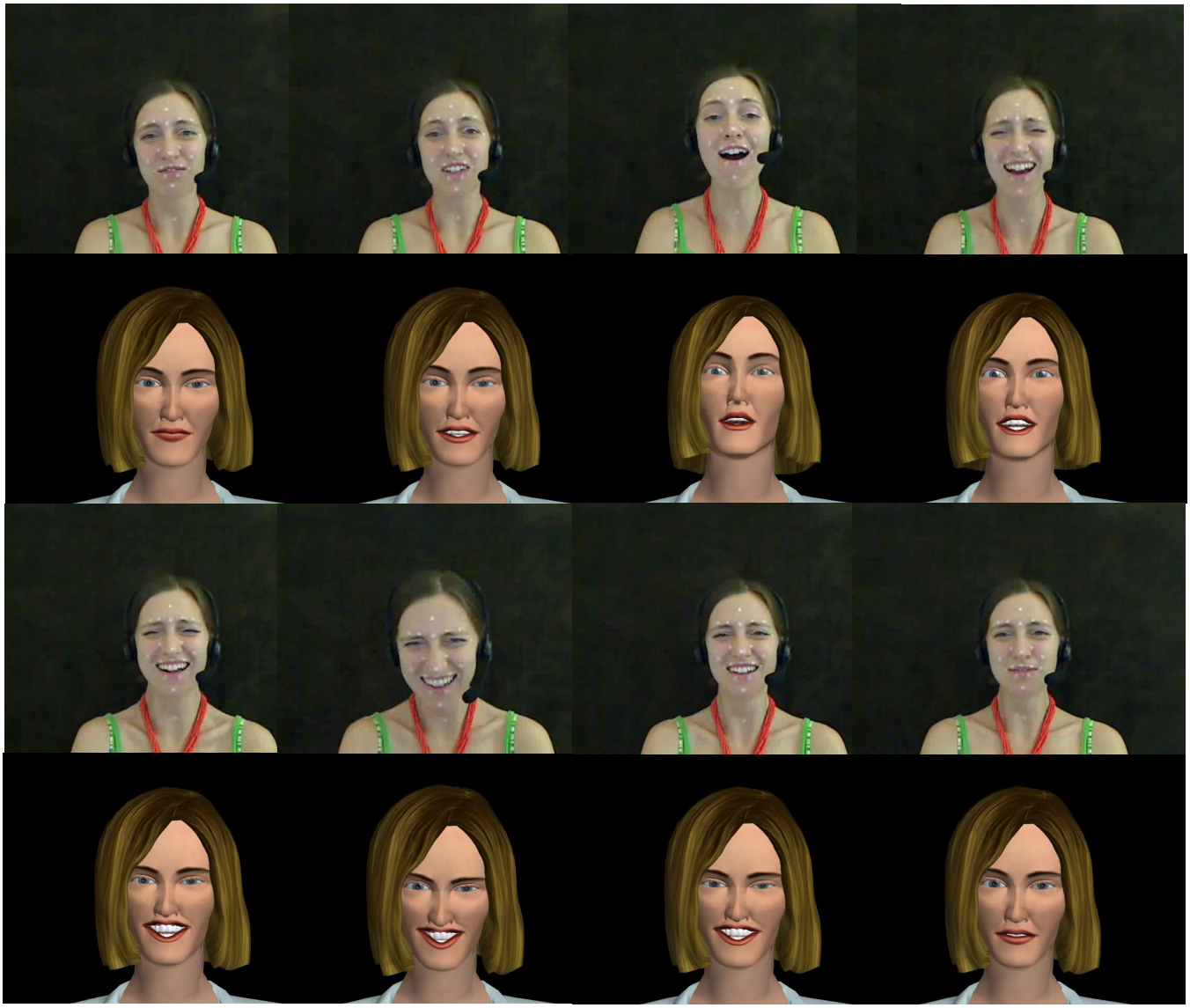

Fig. 13. Animation from Automatic Facial Movements detection

predefined laughs. But the foundation of more complex scripts is ready.

When there are no instructions sent, the real-time 3D rendering component automatically triggers "Idle" animations during which the virtual agent is breathing, making it more realistic and assuring animations continuity.

C. Audiovisual laughter synthesis

In the present work, no new laughter is generated. Instead, existing laughs are re-synthesized. All the animations can thus be prepared. For all the laughter animations, we synthesized separately the acoustic and the visual modalities, using the original audiovisual signals (with synchronized audio and video flows). In consequence the synthesized audio and video modalities are also synchronized. Each acoustic laugh was synthesized and the produced WAVE file was made available to the virtual agent. When the agent receives the instruction to laugh, it loads simultaneously the acoustic (WAVE) file and the BML animation file, and plays them synchronously.

IX. Experiments

A. Participants

Twenty-one participants (13 males; ages ranging from 25 to 56 years, $M = 33.16$, $SD = 8.11$) volunteered to participate. Four participants were assigned to the fixed speech condition, 5 to the fixed laughter condition and 11 to the interactive condition.

B. State and Trait influences on the perception of the virtual agent and its evaluation

Three kinds of subjective ratings were utilized to assess a) habitual and b) actual factors affecting the perception of the virtual agent and c) the evaluation of the interaction. For the habitual factors, two concepts were used: the dispositions towards ridicule and laughter, and the temperamental basis of the sense of humor, with one questionnaire each (PhoPhiKat$< 45 >$; [42]; State-Trait Cheerfulness Inventory,

STCI; [43]). Actual factors were assessed by measuring participant's mood before and after the experiment (state version of the STCI; [44]). The evaluation of the interaction was assessed with the Avatar Interaction Evaluation Form (AIEF; [45]).

1) Habitual Factors:

The assessment of personality variables allowed for a control of habitual factors influencing the perception of the virtual agent, independent of its believability. For example, gelotophobes, individuals with a fear of being laughed at (see [46]), do not perceive any laughter as joyful or relaxing and they fear being laughed at even in ambiguous situations. Therefore, the laughing virtual agent might be interpreted as a threat and the evaluation would be biased by the individuals fear. By assessing the gelotophobic trait, individuals with at least a slight fear of being laughed at can either be excluded from further analysis, or the influence of gelotophobia can be investigated for the dependent variables. Further, the joy of being laughed at (gelotophilia) and the joy of laughing at others (katagelasticism) might alter the experience with the agent, as katagelsticists might enjoy laughing at the agent, while gelotophiles may feel laughed at by the agent and derive pleasures from this. Both dispositions may increase the positive experience of interacting with an agent. The PhoPhiKat-45 is a 45-item measure of gelotophobia ("When they laugh in my presence I get suspicious"), gelotophilia ("When I am with other people, I enjoy making jokes at my own expense to make the others laugh"), and katagelasticism ("I enjoy exposing others and I am happy when they get laughed at"). Answers are given on a 4-point Likert scale ($1 =$ strongly disagree to $4 =$ strongly agree). Ruch and Proyer [42] found high internal consistencies (all alphas $\geq .84$) and high retest-reliabilities $\geq .77$ and $\geq .73$ (three to six months). In the present sample, reliabilities were satisfactory to high and ranged between α = .81 to .83.

Also, it was shown that the traits and states representing the temperamental basis of the sense of humor influence an individual's threshold for smiling and laughter, being amused, appreciating humor or humorous interactions (for an overview see [47]). It was assumed that trait cheerful individuals would enjoy the interaction more than low trait cheerful individuals, as they have a lower threshold for smiling and laughter, those behaviors are more contagious and there are generally more elicitors of amusement to individuals with high scores. For trait bad mood, it was expected that individuals with high scores would experience less positive affect when interacting with the agent, compared to individuals with low scores, as individuals with high scores have an increased threshold for being exhilarated, and they do not easily engage in humorous interactions.

The STCI assesses the temperamental basis of the sense of humor in the three constructs of cheerfulness (CH), seriousness (SE), and bad mood (BM) as both states (STCI-S) and traits (STCI-T). Participants completed the STCI-T before the experiment to be able to investigate the influence of cheerfulness, seriousness and bad mood on the interaction. The standard state form (STCI-S$<30>$; [44]) assesses the respective states of cheerfulness, seriousness and bad mood with ten items each (also on a four-point answering scale). Ruch and Köhler [48] report high internal consistencies for the traits (CH: .93, SE: .88, and BM: .94). The one month test-retest stability was high for the traits (between .77 and .86), but low for the states (between .33 and .36), conforming the nature of enduring traits and transient states.

2) Actual Factors:

Different experiments and studies on the state-trait model of cheerfulness, seriousness, and bad mood showed that participant's mood alters the experience of experimental interventions and natural interactions (for an overview, see [47]). Also, individual's mood changes due to interactions and interventions, for example state seriousness and bad mood decrease when participating carnival celebrations, while cheerfulness increases. Therefore, state cheerfulness, seriousness and bad mood were assessed before and after the experiment to investigate mood influence on the interaction with the agent (with the above mentioned STCI-S).

3) Evaluation:

To evaluate the quality of the interaction with the virtual agent, the naturalness of the virtual agent and cognitions and beliefs toward it, a questionnaire was designed for the purposes of the experiment. The aim of the Avatar Interaction Evaluation Form (AIEF) is to assess the perception of the agent, the emotions experienced in the interaction, as well as opinions and cognitions towards it on broad dimensions. The instrument consists of 32 items and 3 open questions, which were developed following a rational construction approach. The first seven statements refer to general opinions/beliefs and feelings on virtual agents (*e.g.*, "generally I enjoy interacting with virtual agents"). Then, 25 statements are listed to evaluate the experimental session. The following components are included: positive emotional experience (8 items; *e.g.*, "the virtual agent increased my enjoyment"), social (and motivational) aspects (7 items; *e.g.*, "being with the virtual agent just felt like being with another person"), judgment of technical features of the virtual agent/believeability (5 items; *e.g.*, "the laughter of the virtual agent was very natural"), cognitive aspects assigned to the current virtual agent (5 items; *e.g.*, "the virtual agent seemed to have a personality"). All statements are judged on a seven point Likert-scale (1 = strongly disagree to 7 = strongly agree). In the three open questions, participants can express any other thoughts, feelings or opinions they would like to mention, as well as describing what they liked best/least.

4) Further Evaluation Questions and Consent Form:

To end the experimental session, the participants were asked for general questions to assess their liking of candid camera humor in general ("Do you like candid camera-clips in general?" "How funny were the clips?" "How aversive were the clips?" "Would you like to see more clips of this kind?"). All questions were answered on a seven point Likert-scale. Then, participants were asked to give written consent to the use of the collected data for research and demonstration purposes (eNTERFACE workshop and ILHAIRE[8] project).

C. Conditions

1) Overview:

[8]http://www.ilhaire.eu

To create an interaction setting, the participants were asked to watch a film together with the virtual agent. Three conditions were designed (fixed speech, fixed laughter, interactive), systematically altering the degree of expressed appreciation of the clip (amusement) in verbal and non-verbal behavior, as well as different degrees of interaction with the participant's behavior. In the fixed speech and fixed laughter conditions, the agent would be acting independent of the participant, but still be signaling appreciation. In the interactive condition, the agent was responding to the participant's behavior. In other words, only the contextual information was used in the fixed speech and fixed laughter conditions, while the input and decision components (see Sections VI and VII) were active in the interactive condition.

2) Selection of pre-defined time points for the fixed laughter and fixed speech condition:

The pre-defined times were chosen from the stimulus video. Firstly, 14 subjects (three females) watched the video material and annotated the funniness to it on a continuous funniness rating scale (ranging from "not funny at all" to "slightly funny", to "funny", to "really funny" to "irresistibly funny"). Averaged and normalized funniness scores were computed over all subjects, leading to sections with steep increases in funniness (apexes; see Figure 14) over the video. Secondly, the trained raters assigned "punch lines" to the stimulus material, basing on assumptions of incongruity-resolution humor theory. Whenever the incongruous situation/prank was resolved for the subject involved, and amusement in the observer would occur from observing the resolution moment, a peak punch line was assigned. Punch lines were assigned for the first punch line occurring and the last punch line occurring in a given clip. When matching the continuous ratings with the punch lines, it was shown that the funniness apexes did cluster within the first and last punch lines for all subjects and all pranks, apart from one outlier. Table II shows the overall and apex durations of each clip, as well as the number and intensity of the peaks that have been fixed. For the three long apex sections, two responses were fixed, were the averaged funniness ratings peaked. Those peaks were rated on an intensity scale from 1 to 4. Pre-defined time points were controlled for a 1.5s delay in the rating/recording, due to reaction latency of the subjects and motor response delay.

TABLE II
DURATION, APEX AND NUMBER FO FIXED RESPONSES FOR EACH OF THE STIMULUS CLIPS

Clip	Duration (s)	Apex duration (s)	Fixed responses	Intensity
1	95	69	2	4
2	131	56	2	2
3	72	26	1	4
4	72	16	1	3
5	78	50	2	1

Notes: 1. Duration of apex (1st to last punch line). 2. Intensity (1 = strong; 4 = weak)

3) Fixed Speech:

In the fixed speech condition, the agent expressed verbal appreciation in 8 short phrases (*e.g.*, "oh, that is funny", "I liked that one", "ups", "this is great", "how amusing", "phew", nodding, "I wonder what is next") at pre-defined times. The verbal responses were rated for intensity on a four point scale and matched to the intensity scores of the pre-defined time points.

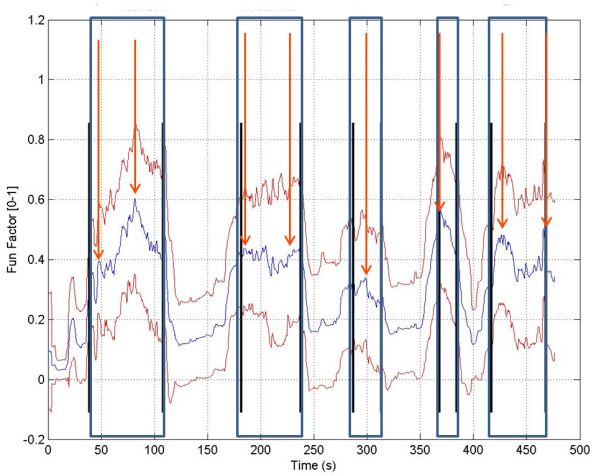

Fig. 14. Continuous funniness ratings (means in blue and standard deviations in red) over the stimulus video for 14 subjects and expert assigned punch lines (first and last, in blue) to each clip. Red arrows indicate time points for fixed responses.

4) Fixed Laughter:

In the fixed laughter condition, the agent laughed at pre-defined times during the video. The times were the same as the time points in the fixed speech condition. The agent displayed 8 laughs which varied in intensity and duration, according to the intensity ratings of the pre-defined time points. A laughter bout may be segmented into an onset (*i.e.*, the pre-vocal facial part), an apex (*i.e.*, the period where vocalization or forced exhalation occurs), and an offset (*i.e.*, a post-vocalization part; often a long-lasting smile fading out smoothly; see [21]). Therefore, the onset was emulated by an independent smiling action just before the laughter (apex) would occur at the fixed time. The offset of the laughter was already integrated in the 8 laughter chosen.

5) Interactive Condition:

In the interactive condition, which follows the architecture presented in Section V and Figure 1, the agent was using two sources of information to respond to the participant: the continuous funniness ratings to the clip (context, shown in Figure 14) and the participant's acoustic laughter vocalizations. The dialog manager was receiving these two information flows and continuously taking decisions about whether and how the virtual agent had to laugh, providing intensity and duration values of the laugh to display. These instructions were then transmitted to the audiovisual synthesis modules. Due to the limited number of laughs available for synthesis (14 at the time of the experiments), it was decided to cluster them into 4 groups based on their intensities and durations. The output of the dialog manager is then pointing to one of the clusters, inside which the laugh to synthesize is randomly picked.

D. Problems encountered

Several problems appeared during the experiments.

First of all, the computers we used were not powerful enough to run all the components on a single computer. We had to connect four computers together: one master computer running the stimulus video and the Kinect recording and analysis (+ the context), one computer running a webcam with shoulder movement tracking driven by Eyesweb, another one running the dialog manager and finally one computer for displaying the virtual agent. Still, the master computer could not record the video stream from the Kinect. We decided to run the experiments without recording that video as we still have the webcam recording, but this issue should be investigated in the future. Furthermore, during some experiments, data transmission from one computer to the other was suffering from important delays (5-10s), which obviously affect the quality of the interaction. Reducing these delays will be one of the most important future developments.

Second, the audio detection module had been trained with data containing mostly laughs, and relatively few other noises. Hence, there was confusion between laughter and other loud noises. In addition, the detection was audio-only, which does not enable to take smiles or very subtle laughs (with low audio) into account. We are already working on improving the laughter detection and including other modalities (video, respiration) to increase its robustness.

Third, from the training data, it appeared that the context was by far the best factor to explain participants' laughs: in consequence, the dialog manager did not pay attention to what the participant was doing, but only triggered laughs from the contextual input. Since this is undesirable behavior in the interactive condition (which is in that case actually similar to the fixed laughter condition, as every reaction is only context-dependent), we decided to omit the context in the interactive condition: the virtual agent was then only reacting to what the participant was doing. Better models should be built in the future to allow both context and participant's reactions to be considered simultaneously.

Fourth, the pool of available laughs for synthesis is currently limited. There are not a lot of laughs from one single voice for which we have good quality synthesis for both the audio and the visual modalities. This limits the range of actions the virtual agent is able to perform and some participants with whom the agent laughed a lot might have noticed some repetitions. This will be improved in the future with 2 solutions: 1) a larger pool of available laughs 2) the possibility to generate new laughter transcription and/or modify existing ones in real-time.

Finally, a connection problem with the respiration sensor prevented us from recording respiration data.

E. Procedure of the evaluation study

Participants were recruited through e-mail announcement of an "evaluation study of the Laugh Machine project" at the eNTERFACE workshop. As an incentive, participants were offered a feedback on the questionnaire measures on request. It was announced that the study consisted of the filling in of questionnaires (approximately 30-45 minutes) and a session of 30 minutes on two given days. No further information on the aims of the study was given. Participants chose a date for the experimental session via the Internet and received confirmation by email.

At the experimental session, participants were welcomed by one of the two female experimenters and asked to fill in the STCI and the PhoPhiKat. Then, participants were asked to fill in the STCI-S to assess their current mood. Meanwhile, the participants were assigned to one of the three conditions. Afterwards, the second female experimenter accompanied the participant to the experimenting room, where the participant was asked to sit in front of a television screen. A camera allowed for the frontal filming of the head and shoulder, as well as upper body of the participant. Two male experimenters concerned with the technical components were present. Participants were asked for consent to have their shoulder and body movements recorded. They were also given headphones to hear the virtual agent. The experimenter explained that the participant was asked to watch a film together with Poppy and that the experimenters would leave the room when the experiment started. Once the experimenters left the room, the agent did greet the participant ("Hi, I'm Greta. I'm looking forward to watch this video with you. Let's start") and subsequently, the video started. After the film, the experimenters entered the room again and the female experimenter accompanied the participant back to the location where the post measure of the STCI-S, as well as the AIEF and five further evaluation questions were filled in. After all questionnaires were completed, the first female experimenter debriefed the participant and asked for written permission to use the obtained data.

The following setup was used in this experimental session (see Figure 15). Two LCD displays were used: the bigger one (46") was used to display the stimuli (the funny film, see Section III). The smaller (19") LCD display placed on the right side of the big one was used to display the agent (a close-up view of the agent with only the face visible was used). Four computers were used to collect the user data, run the Dialog Module and to control the agent audio-visual synthesis. Participant's behaviors were collected through a Kinect (sound, depth map, and camera) and a second webcam synchronized with the EyesWeb software (see Section VI). Because of technical issues we were not able to use the respiration sensor in this experimental session. Participants were asked to sit on a cushion about 1m from the screen. They were asked to wear headphones.

In the evaluation we have used 14 laugh episodes from the AVLC dataset (subject 5). For consistency reasons we have used only one female agent (*i.e.*, Poppy) and the animation created with only one method *i.e.* automatic facial movements detection (see Section VIII-B3).

X. Results

A. Preliminary Analysis

Scale means for cheerfulness, seriousness, bad mood, gelotophobia, gelotophilia and katagelasticism were investigated. The sample characteristics of the PhoPhiKat and the STCI-T

Fig. 15. Setup of the experiment.

resembled norm scores for adult populations. In this sample, the internal consistencies were satisfactory for all trait scales, ranging from $\alpha = .74$ for trait seriousness, to $\alpha = .91$ for trait cheerfulness. In respect to trait variables biasing the evaluation, three subjects were identified for exceeding the cut-off point for gelotophobia. Means for the state cheerfulness, seriousness and bad mood scores showed higher state bad mood scores before the experiment, compared to previous participants of studies on personality and humor. In respect to the AIEF, the internal consistencies (Cronbach's alpha) of the scales were satisfactory, ranging from $\alpha = .78$ (cognitive aspects) to $\alpha = .90$ (positive emotional experience).

B. Traits

In line with previous findings, trait cheerfulness was correlated negatively to trait bad mood ($r = -.61, p < .01$), as well as trait seriousness ($r = -.16, n.s.$), but less strongly to the latter one. Trait seriousness and bad mood were correlated positively ($r = .22, n.s.$). Gelotophobia was correlated negatively to gelotophilia ($r = -.50, p < .05$), as well as (but less so) to katagelasticism. The latter negative (but not significant; $r = -.35, p = .117$) correlation was unusual, as gelotophobia usually shows zero correlations to katagelasticism. Katagelasticism was positively related to gelotophilia ($r = .26, n.s.$). Generally, correlations of the AIEF to the trait scale did not reach statistical significance. Correlating the dimensions and items of the AIEF to gelotophobia (bivariate Pearson correlations) showed, that three of the four AIEF scales were negatively correlated with gelotophobia, indicating that higher scores in gelotophobia went along with less positive emotions, less assignment of cognition and less believability of the virtual agent to participants with higher scores. Feeling social presence by the agent was positively correlated to gelotophobia. Gelotophilia correlated positively with all dimensions of the AIEF. Further, higher scores in katagelasticism went along with more positive emotions, higher perceived believability and higher perceived social presence. With regard to the temperamental basis of the sense of humor, the highest correlations to the AIEF dimensions were found for trait bad mood. Unlike a priori assumptions, trait bad mood correlated positively to the AIEF dimensions and correlations to trait cheerfulness were generally very low. Trait seriousness was correlated negatively to the AIEF scales.

C. States

Correlating the states to their respective traits showed that trait cheerfulness was positively correlated to state cheerfulness both pre and post the experiment (but all $n.s.$). Trait seriousness was positively correlated to seriousness after the experiment, whereas trait bad mood was negatively correlated to both, bad mood pre and post the experiment (both $p < .01$). In this sample, a few individuals with low scores in trait bad mood came to the experiment with high values in state bad mood, whereas a few individuals with high scores on bad mood came to the experiment with comparably low scores in state bad mood. Descriptive analysis of the mood before and after the experiment, it was found that the interaction with the virtual agent led to a decrease in seriousness over all conditions, whereas state cheerfulness stayed stable. State bad mood before the experiment predicted lower scores on the AIEF dimensions, suggesting that individuals that feel more grumpy or sad generally experience less positive emotions with the virtual agent, assign the agent less cognitive capability, experience less social interaction and judge it as less believable (all $r < -.489$, $p < .05$).

D. AIEF Scales/Dimensions

Due to the low cell sizes, no test of significance could be performed to testing the influence of the condition on the AIEF dimensions. Nevertheless descriptive inspection of the group means showed that the conditions differed in their elicitation of positive outcomes on all dimensions of the AIEF. The interactive condition yielded highest means on all four dimensions, implying that the participants felt more positive emotions, felt more social interaction with the agent, considered it more natural and assigned it more cognitive capabilities than in the fixed conditions (see Figure 16). The means of the interactive condition were followed by the means of the fixed laughter condition.

Interestingly, the fixed speech condition yielded similarly high scores on the beliefs on cognition as the interactive condition, whereas the other means were numerically lowest for the fixed speech condition.

The results stayed stable when excluding the three individuals exceeding the cut-off point for gelotophobia.

E. Open Answers

Out of the 21 participants, 14 gave answers to the question of what they liked least about the session. Half of the participants mentioned that the video was not very funny or would have been funnier with sound. Two participants mentioned that they could not concentrate on both, the virtual agent and the film. Two of the three gelotophobes gave feedback (subject 2: "Poppy's expression while laughing was more a smirk than a

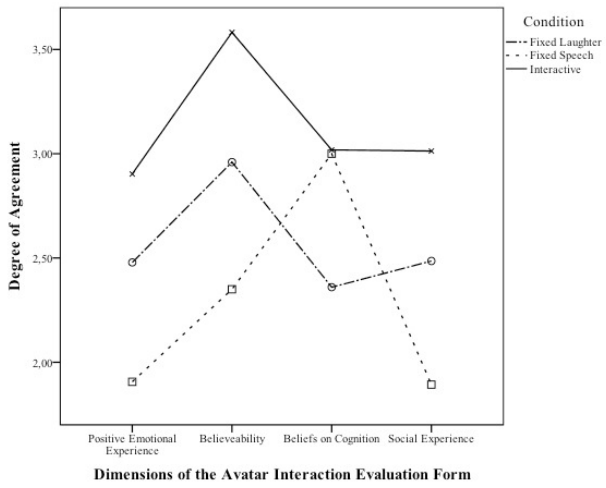

Fig. 16. Profiles of the means in the AIEF scales for the three experimental conditions separately

laugh"; subject 21: "it's hard to act naturally when watching a film when you feel like you should laugh"). Seventeen participants responded to what was liked best about the session. Best liked was the laughter of the virtual agent through the headphones (it was considered amusing and contagious; three nominations), the video (five nominations), the set up (four nominations) and one participant stated: "It was interesting to see in what situations and in what manner the virtual agent responded to my laughter and to funny situations respectively" (subject 12).

XI. Collected Data

The multimodal laughter corpora of human to human interactions are rare. Even more seldom are corpora of human-machine interaction that contain any episodes of laughter. The evaluation of our interactive laughter system gave us the unique opportunity to gather new data about the human behavior in such human-machine interactive scenario. Consequently, we have collected multimodal data from participants to our experiments. In more details our corpus contains:

- audio data, recorded by the Kinect at 16kHz and stored in mono WAVE files, PCM 16bits
- Kinect depth maps
- two web cameras
- data on the shoulders movement extrapolated from the video stream (for this purpose two small markers were placed on the shoulders of each participant)

All these data can be synchronized with the context (see Section IX-C2) and the agent reactions. The collected corpus is an important result of the Laugh Machine Project. It will be widely used in the ILHAIRE project and will become freely available for the research purposes.

XII. Conclusions

The first results of the evaluation experiment are highly promising: it was shown that the three conditions elicited different degrees of positive emotions in the participants, the amount of social interaction induced, as well as the cognitions and capability assigned to the agent. Also, the believability differed for all three conditions. It was shown that the interactive condition yielded the most positive outcomes on all dimensions, implying that the feedback given to the participant by mimicking his or her laughter is best capable of creating a "mutual film watching experience" that is pleasurable.

In sum, expressing laughter increases the positive experience in the interaction with an agent, when watching an amusing video (both laughter conditions elicited more positive emotions), compared to the fixed speech condition. The fixed speech condition yielded numerically lowest means on the AIEF dimensions, apart from the dimension "beliefs on cognition", where the means where as high as in the interactive condition, implying that speech leads to the assignment of cognitive ability equally as much as responding to the participant's behavior. Naturally, the fixed speech conditions should yield the lowest scores, as there was no laughter expressed in this conditions and some items targeted the contagiousness and appropriateness of the laughter displayed by the agent.

Obviously, in the interactive condition, the amount of laughter displayed by the agent varied for each participant, depending on how many times the participants actually laughed. Therefore, the agent behaved similar to the participant, which seems to be natural and comfortable for the participant. Nevertheless, the current state of data analysis does not allow to differentiating between individuals who displayed a lot of laughter—and consequently had a lot of laughter feedback by the agent—and individuals who showed only little laughter—and received little laughter feedback by the agent. An in depth analysis of the video material obtained during the eNTERFACE evaluation experiment will allow for an investigation of how many times the participants actually laughed and how this influenced the perception of the setting. This will be done by applying the FACS [39]. Further, an analysis of the eye movements (gaze behavior) will allow for an estimation of the attention paid to the agent.

The results of the trait and state cheerfulness, seriousness, and bad mood variables clearly show the importance of including personality variables into such evaluation experiments. Especially state bad mood influenced the interaction and latter perception of the virtual agent, leading to a mood dependent bias. Individuals with high scores in state bad mood before the experiment evaluated the virtual agent less favorably. This is likely due to their enhanced threshold for engaging in cheerful/humorous situations/interactions and—in the case of grumpiness—their unwillingness to be exhilarated and—in the case of depressed/melancholic mood—the incapability to be exhilarated. Therefore, personality should always be controlled for in future studies. Generally, there was sufficient variance in the gelotophobia scores, even in the little sample obtained in the evaluation. Gelotophobia showed some systematic relations to the dimensions of the AIEF. For future studies, the assessment and control of gelotophobia is essential to get unbiased evaluations of an agent. Furthermore, those results might help the understanding of the fear of being laughed at and how it influences the thoughts, feelings and behavior of

individuals with gelotophobia.

Nevertheless, more participants are needed to test the hypothesis on the influence of the condition on the AIEF dimensions in order for any statistically significant differences between the conditions to be found. To improve the experimental set up, challenges from eNTERFACE, as well as the participant's feedback will be integrated to optimize the design and procedure. For example, the stimulus video consisted of only one type of humorous material. It is well established in psychological research that inter-individual differences exist in the appreciation of types of humor. Therefore, a lack of experienced amusement on the side of the participant might also be due to the disliking of candid camera clips, as one specific type of humor. Any manipulation by the experimental conditions should not be overshadowed by the quality/type of stimulus video. Therefore, a more representative set of clips with sound is needed (presented in counter-balanced order, also extending the overall interaction time with the virtual agent).

Furthermore, it needs to be clear to participants beforehand, what the virtual agent is capable of doing. In the beginning of the experiment, the virtual agent should display some laughter, so the participant knows, that the virtual agent would be capable of showing this behavior. This ensures, that the participant will not be solely surprised and amused by the fact, the virtual agent can laugh, when it eventually does during the course of the film. If this information is not available to participants, it might be that the amusement is only due to the excitement/pleasure of the technical development of making a virtual agent laugh. Ruch and Ekman's [21] overview on the knowledge about laughter (respiration, vocalization, facial action, body movement) illustrated the mechanisms of laughter, and defined its elements. While acknowledging that more variants of this vocal expressive-communicative signal might exist, they focused on the common denominators of all forms but proposed distinguishing between laughing spontaneously (emotional laughter) and laughing voluntarily (contrived or faked laughter). In this experiment, only displays of amusement laughter (differing in intensity and duration) were utilized. Further studies may also include different variants of types of laughter.

On the technical side, the biggest outcome of the project is a full processing chain with components that can communicate together to perform multimodal data acquisition, real-time laughter-related analysis, output laughter decision and audio-visual laughter synthesis. Progresses have been accomplished on all these aspects during the Laugh Machine project. We can cite the development of the respiration sensor and the integration of all input devices in a synchronized framework, which will enable multimodal laughter detection; the construction of a real-time, speaker independent, laughter intensity estimator; the design of the first dialog manager dealing with laughter; the first advances in acoustic laughter synthesis with the introduction of HMM-based processes; the four different animation techniques that have been implemented; or the unique database of humans interacting with a laughing virtual agent that has been collected.

Each of these components can be improved and several issues arose during the experiments. Without going into details for each of the involved components, future works will include: improving the laughter detection and intensity computation with the help of visual and respiration signals; reducing the communication delays between the computers hosting the different modules; better balancing the influence of the context in the dialog manager; extending the range of output laughs by allowing laughs to be generated or modified on the fly; ensuring that all experimental data can be recorded flawlessly; or adapting the virtual agent's behavior to the participant's personality (*e.g.*, gelotophobe) and mood to maximize the participant's perception of the interaction. Also, future agents may not only include facial expressions and vocal utterances, as laughter also entails lacrimation, respiration, body movements (*e.g.*, [49]), body posture and vocalization.

However, despite all the identified issues, the first evaluation results are positive. This is very encouraging and indicates that the full LaughMachine system, while imperfect, is already working and providing us with both a nice benchmark and a reusable framework to evaluate future developments.

ACKNOWLEDGMENT

This work was supported by the European FP7-ICT projects ILHAIRE (FET, grant $n°270780$), CEEDS (FET, grant $n°258749$) and TARDIS (STREP, grant $n°288578$). The authors would also like to thank the organizers of the eNTERFACE'12 Workshop for making the project possible and making available high quality material and recording rooms to us. Finally, all the participants of our experiments are gratefully acknowledged.

REFERENCES

[1] L. Kennedy and D. Ellis, "Laughter detection in meetings," in *NIST ICASSP 2004 Meeting Recognition Workshop*, Montreal, May 2004, pp. 118–121.

[2] K. P. Truong and D. A. van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Communication*, vol. 49, pp. 144–158, 2007.

[3] M. T. Knox and N. Mirghafori, "Automatic laughter detection using neural networks," in *Proceedings of Interspeech 2007*, Antwerp, Belgium, August 2007, pp. 2973–2976.

[4] S. Petridis and M. Pantic, "Fusion of audio and visual cues for laughter detection," in *Proceedings of the 2008 international conference on Content-based image and video retrieval*. ACM, 2008, pp. 329–338.

[5] ——, "Audiovisual discrimination between laughter and speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, 2008, pp. 5117–5120.

[6] ——, "Audiovisual laughter detection based on temporal features," in *Proceedings of the 10th international conference on Multimodal interfaces*. ACM, 2008, pp. 37–44.

[7] S. Petridis, A. Asghar, and M. Pantic, "Classifying laughter and speech using audio-visual feature prediction," in *Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. Dallas, Texas: IEEE, 2010, pp. 5254–5257.

[8] S. Sundaram and S. Narayanan, "Automatic acoustic synthesis of human-like laughter," in *Journal of the Acoustical Society of America*, vol. 121, no. 1, January 2007, pp. 527–535.

[9] E. Lasarcyk and J. Trouvain, "Imitating conversational laughter with an articulatory speech synthesis," in *Proceedings of the Interdisciplinary Workshop on The Phonetics of Laughter*, August 2007, pp. 43–48.

[10] T. Cox, "Laughter's secrets: faking it – the results," *New Scientist*, 27 July 2010. [Online]. Available: http://www.newscientist.com/article/dn19227-laughters-secrets-faking-it--the-results.html

[11] P. DiLorenzo, V. Zordan, and B. Sanders, "Laughing out loud: control for modeling anatomically inspired laughter using audio," in *ACM Transactions on Graphics (TOG)*, vol. 27, no. 5. ACM, 2008, p. 125.

[12] D. Cosker and J. Edge, "Laughing, crying, sneezing and yawning: Automatic voice driven animation of non-speech articulations," in *Proc. of Computer Animation and Social Agents (CASA09)*. Citeseer, 2009, pp. 21–24.

[13] J. Urbain, R. Niewiadomski, E. Bevacqua, T. Dutoit, A. Moinet, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner, "AVLaughterCycle: Enabling a virtual agent to join in laughing with a conversational partner using a similarity-driven audiovisual laughter animation," *Journal on Multimodal User Interfaces*, vol. 4, no. 1, pp. 47–58, 2010.

[14] S. Shahid, E. Krahmer, M. Swerts, W. Melder, and M. Neerincx, "Exploring social and temporal dimensions of emotion induction using an adaptive affective mirror," in *27th international conference extended abstracts on Human factors in computing systems*. ACM, 2009, pp. 3727–3732.

[15] S. Fukushima, Y. Hashimoto, T. Nozawa, and H. Kajimoto, "Laugh enhancer using laugh track synchronized with the user's laugh motion," in *Proceedings of the 28th of the international conference on Human factors in computing systems (CHI'10)*, 2010, pp. 3613–3618.

[16] C. Becker-Asano, T. Kanda, C. Ishi, and H. Ishiguro, "How about laughter? Perceived naturalness of two laughing humanoid robots," in *Affective Computing and Intelligent Interaction*, 2009, pp. 49–54.

[17] C. Becker-Asano and H. Ishiguro, "Laughter in social robotics - no laughing matter," in *Intl. Workshop on Social Intelligence Design (SID2009)*, 2009, pp. 287–300.

[18] G. Mckeown, M. F. Valstar, R. Cowie, and M. Pantic, "The semaine comary:2007kkrpus of emotionally coloured character interactions," in *Proceedings of IEEE Int'l Conf. Multimedia, Expo (ICME'10), Singapore*, July 2010, pp. 1079–1084.

[19] J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner, "The AVLaughterCycle database," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010.

[20] J. Urbain and T. Dutoit, "A phonetic analysis of natural laughter, for use in automatic laughter processing systems," in *International Conference on Affective Computing and Intelligent Interaction (ACII2011)*, Memphis, Tennesse, October 2011, pp. 397–406.

[21] W. Ruch and P. Ekman, "The expressive pattern of laughter," in *Emotion, qualia and consciousness*, A. Kaszniak, Ed. Tokyo: World Scientific Publishers, 2001, pp. 426–443.

[22] The Apache Software Foundation, "Apache ActiveMQ™ [computer program webpage]," http://activemq.apache.org/, consulted on August 24, 2012.

[23] J. Wagner, F. Lingenfelser, and E. André, "The social signal interpretation framework (SSI) for real time signal processing and recognition," in *Proceedings of Interspeech 2011*, 2011.

[24] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[25] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*, ser. MM '10. New York, NY, USA: ACM, 2010, pp. 1459–1462.

[26] R. Niewiadomski, J. Urbain, C. Pelachaud, and T. Dutoit, "Finding out the audio and visual features that influence the perception of laughter intensity and differ in inhalation and exhalation phases," in *Proceedings of the ES 2012 4th International Workshop on Corpora for Research on EMOTION SENTIMENT & SOCIAL SIGNALS, Satellite of LREC 2012*, Istanbul, Turkey, May 2012.

[27] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[28] M. Filippelli, R. Pellegrino, I. Iandelli, G. Misuri, J. Rodarte, R. Duranti, V. Brusasco, and G. Scano, "Respiratory dynamics during laughter," *Journal of Applied Physiology*, vol. 90, no. 4, p. 1441, 2001.

[29] A. Feleky, "The influence of the emotions on respiration." *Journal of Experimental Psychology*, vol. 1, no. 3, pp. 218–241, 1916.

[30] A. Camurri, P. Coletta, G. Varni, and S. Ghisio, "Developing multimodal interactive systems with eyesweb xmi," in *Proceedings of the 2007 Conference on New Interfaces for Musical Expression (NIME07)*, 2007, p. 302305.

[31] B. Lukas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th international joint conference on Artificial intelligence*, 1981.

[32] D. Winter, "Biomechanics and motor control of human movement," 1990.

[33] W. Sethares and T. Staley, "Periodicity transforms," *Signal Processing, IEEE Transactions*, vol. 47, no. 11, pp. 2953–2964, 1999.

[34] K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to English," in *IEEE Speech Synthesis Workshop*, Santa Monica, California, September 2002, pp. 227–230.

[35] K. Oura, "HMM-based speech synthesis system (HTS)," http://hts.sp.nitech.ac.jp/, consulted on June 22, 2011.

[36] T. Yoshimura, "Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for hmm-based text-to-speech systems," Ph.D. dissertation, Ph. D. thesis, Nagoya Institute of Technology, 2002.

[37] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, pp. 968–981, 2012.

[38] H. Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[39] P. Ekman, W. Friesen, and J. Hager, "Facial action coding system: A technique for the measurement of facial movement," 2002.

[40] J. Saragih, S. Lucey, and J. Cohn, "Deformable model fitting by regularized landmark mean-shift," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2011.

[41] M. Zeiler, G. Taylor, L. Sigal, I. Matthews, and R. Fergus, "Facial expression transfer with input-output temporal restricted boltzmann machines," in *Neural Information Processing Systems Conference NIPS 2011*, 2011, pp. 1629–1637.

[42] W. Ruch and R. Proyer, "Extending the study of gelotophobia: On gelotophiles and katagelasticists," *Humor: International Journal of Humor Research*, vol. 22, no. 1-2, pp. 183–212, 2009.

[43] W. Ruch, G. Köhler, and C. Van Thriel, "Assessing the "humorous temperament": Construction of the facet and standard trait forms of the state-trait-cherrfulness-inventory–STCI." *Humor: International Journal of Humor Research; Humor: International Journal of Humor Research*, vol. 9, pp. 303–339, 1996.

[44] W. Ruch, G. Köhler, and C. Van Thriel, "To be in good or bad humour: Construction of the state form of the state-trait-cheerfulness-inventory–STCI," *Personality and Individual Differences*, vol. 22, no. 4, pp. 477–491, 1997.

[45] J. Hofmann, T. Platt, and W. Ruch, "Avatar interaction evaluation form (AIEF)," 2012, unpublished research instrument.

[46] W. Ruch and R. Proyer, "The fear of being laughed at: Individual and group differences in gelotophobia." *Humor: International Journal of Humor Research*, vol. 21, pp. 47–67, 2008.

[47] W. Ruch and J. Hofmann, "A temperament approach to humor," in *Humor and health promotion*, P. Gremigni, Ed. New York: Nova Science Publishers, 2012.

[48] W. Ruch and G. Köhler, "A temperament approach to humor," in *The sense of humor: Explorations of a personality characteristic*, W. Ruch, Ed. Berlin: Mouton de Gruyter, 2007, pp. 203–230.

[49] G. Hall and A. Allińn, "The psychology of tickling, laughing, and the comic," *The American Journal of Psychology*, vol. 9, no. 1, pp. 1–41, 1897.